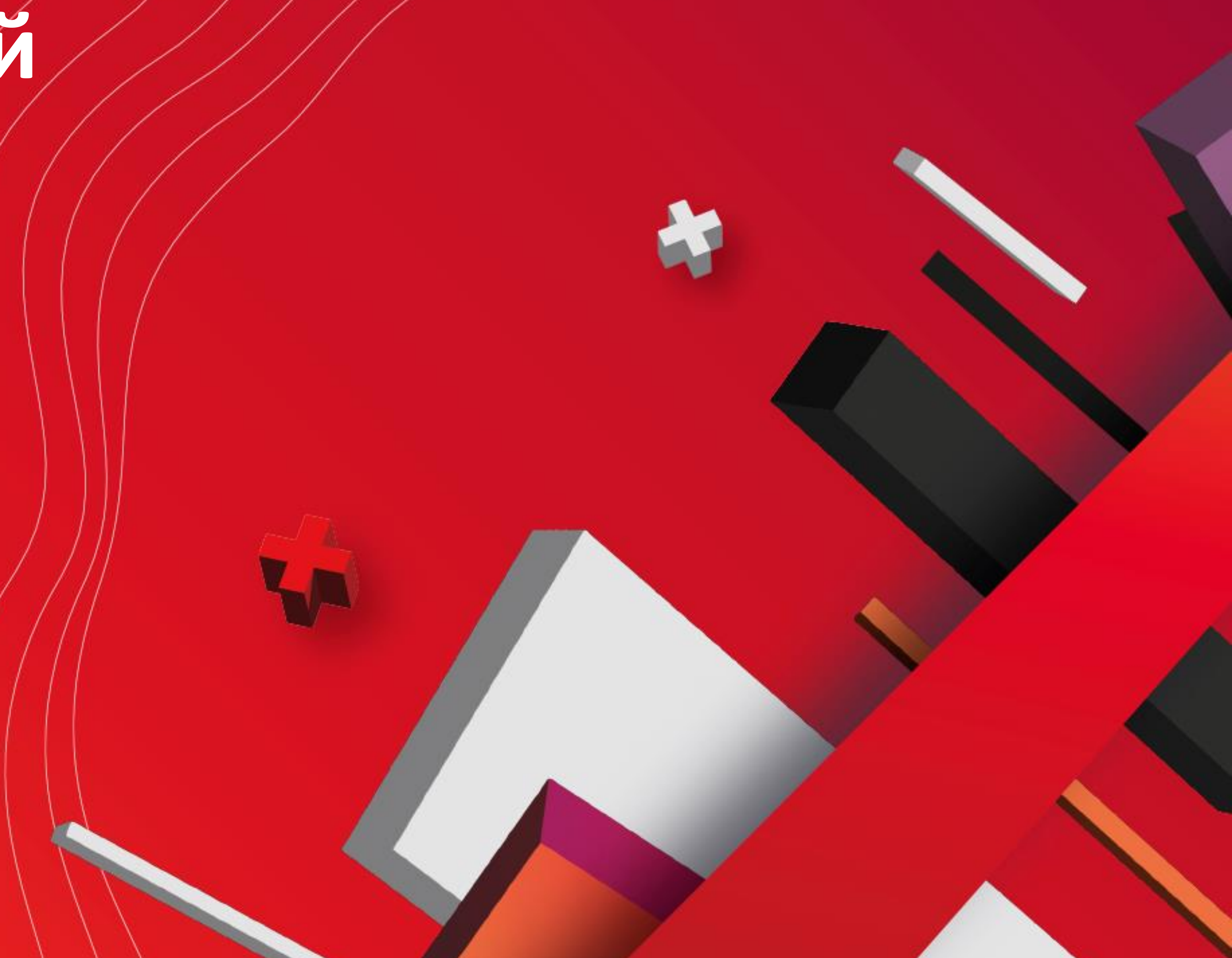


Наша Машина баз данных (как Oracle Exadata, только про PostgreSQL) и система управления к ней

Константин Аристов,
Скала^р, техлид





Скала^р сегодня:

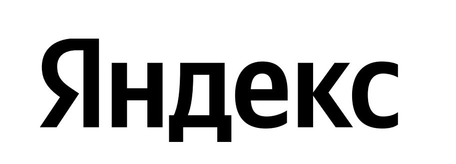


разработка и производство модульной платформы для высоконагруженных государственных и корпоративных информационных систем

6 лет
серийного выпуска

150+
комплексов в промышленной эксплуатации

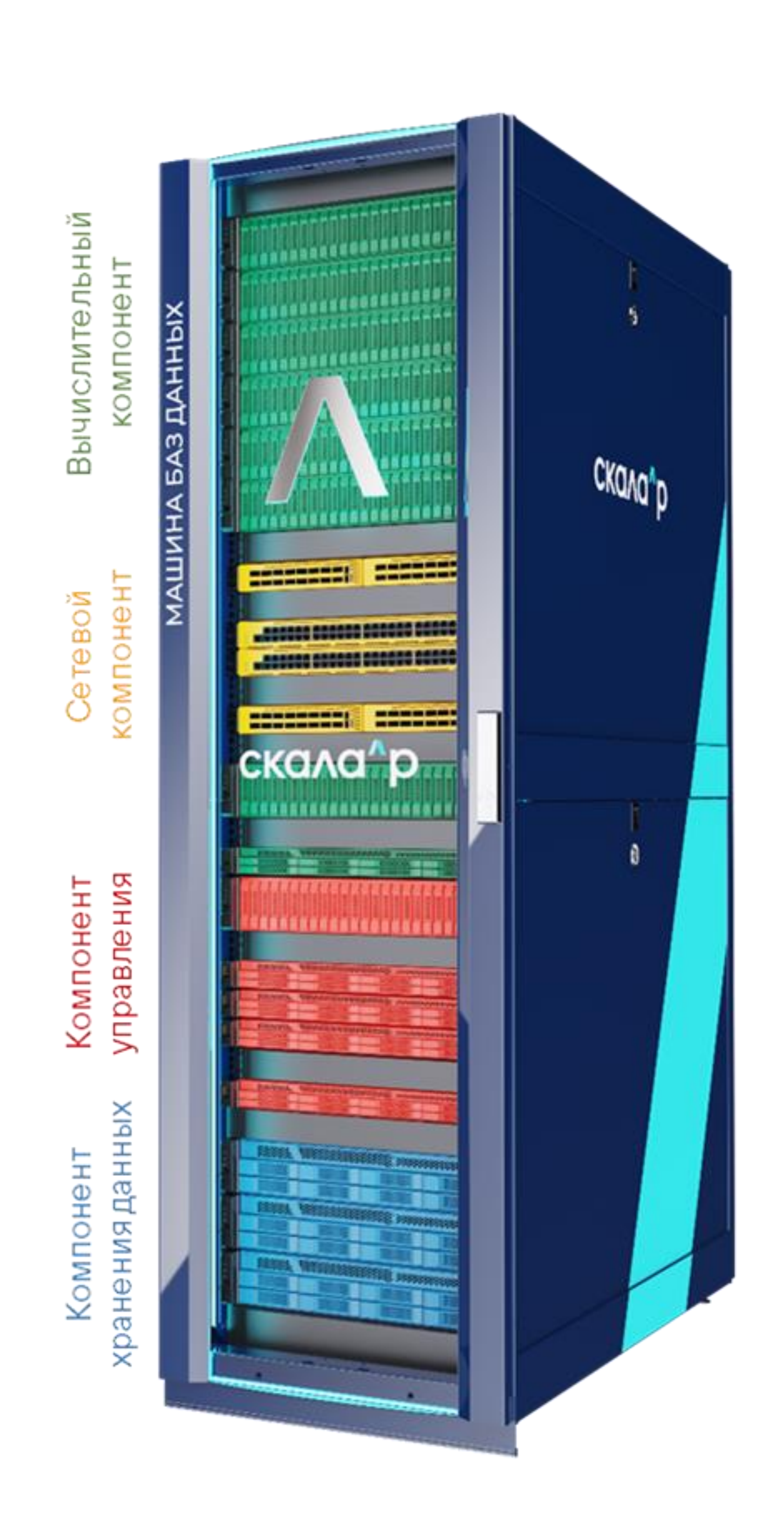
2500+
вычислительных узлов

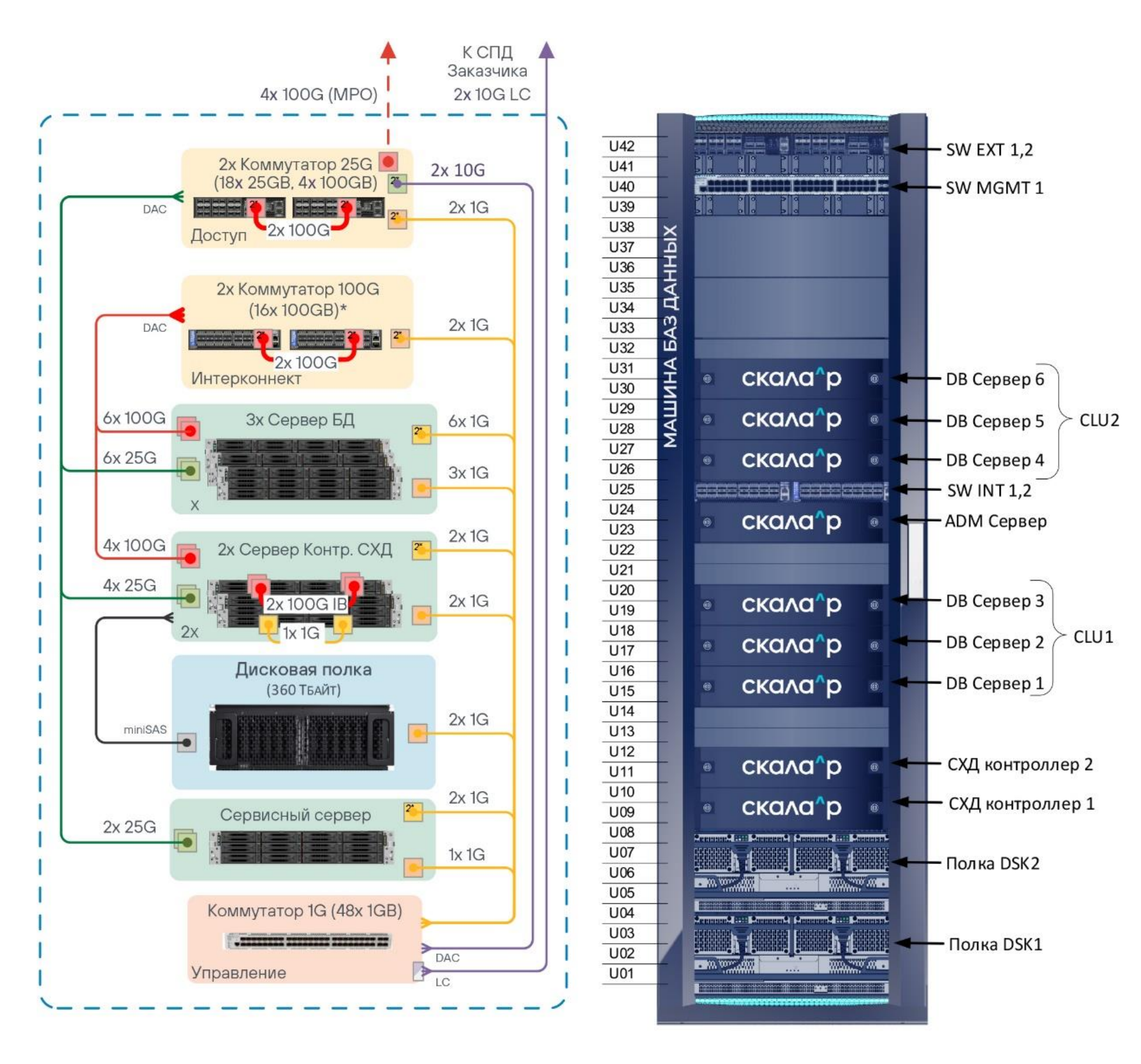




Снаружи и внутри

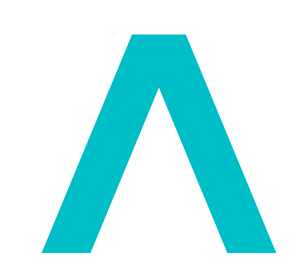








Снаружи и внутри





Сетевой компонент:

- Сетевой узел доступа (36 портов 10|25 ГБ/с + 4 порта 100 ГБ/с)
- Сетевой узел интерконнекта (28 портов 100 ГБ/с)

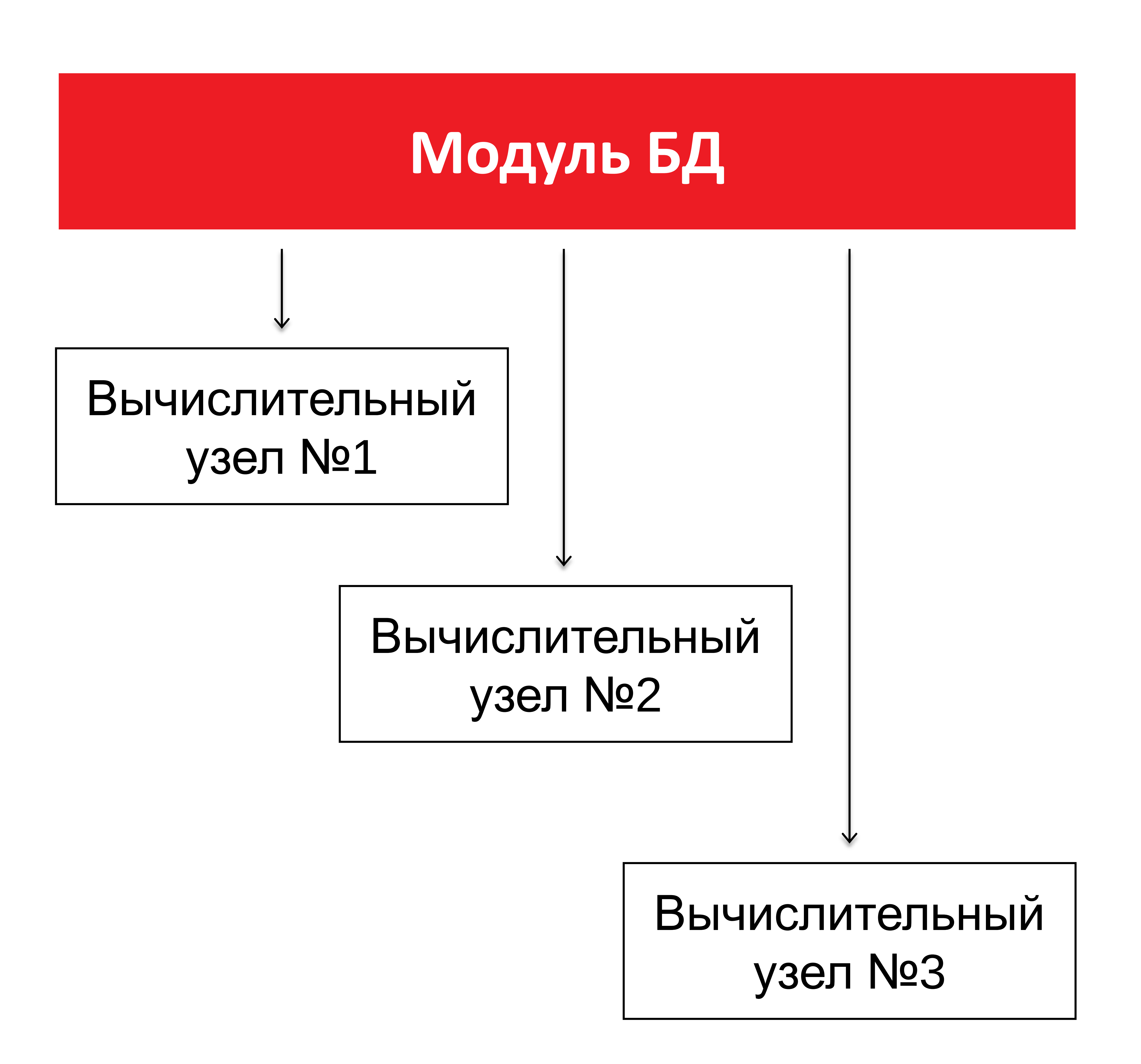
Компонент управления:

- Служебный узел
- Сетевой узел управления



Снаружи и внутри

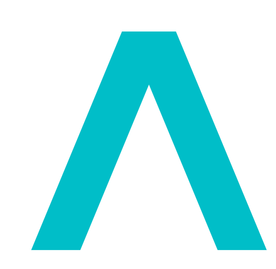




3 вычислительных узла, в каждом:

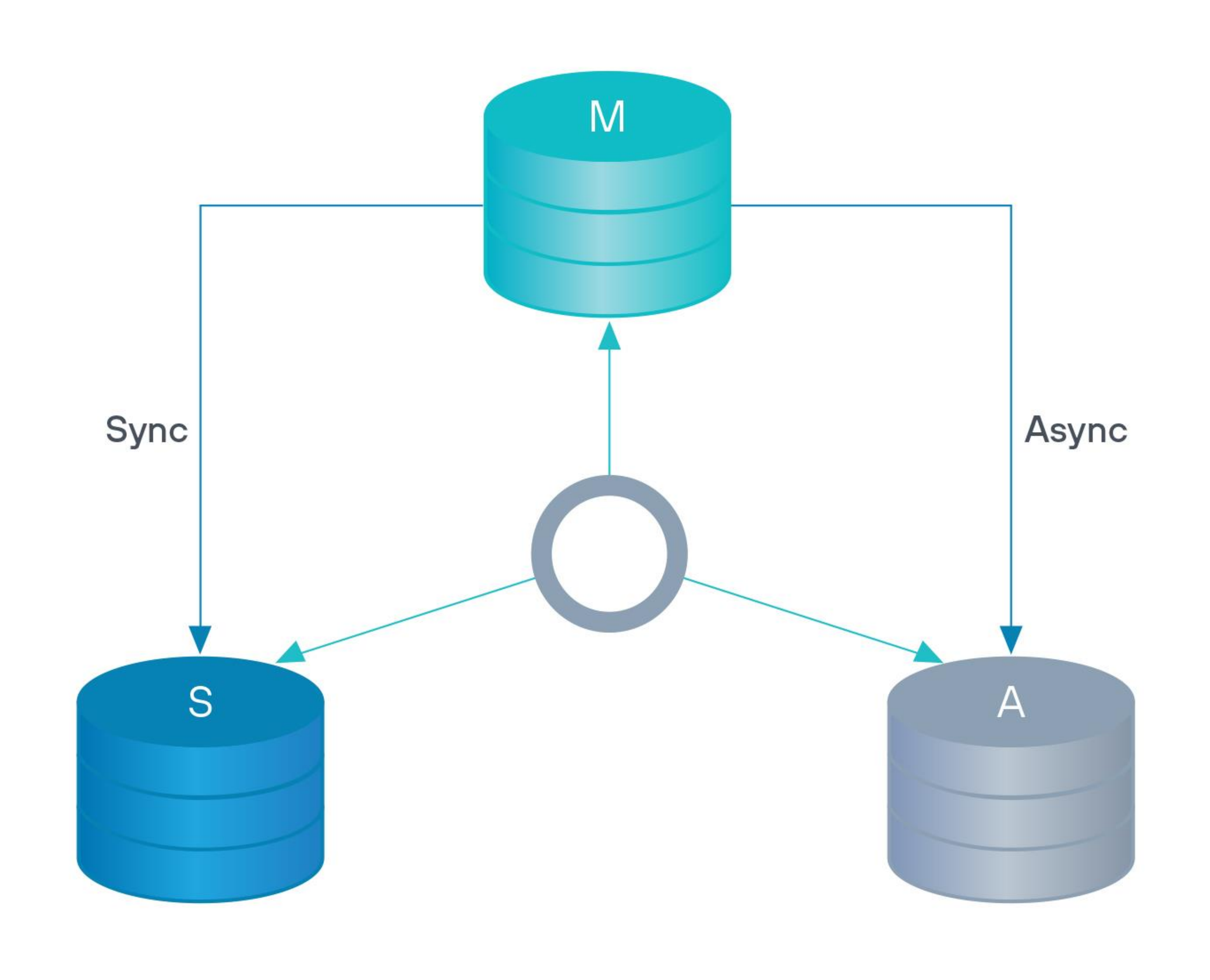
- 48 ядер 3.0 ГГЦ
- 768 ГБ оперативной памяти
- 10-20 ТБ дискового пространства для БД

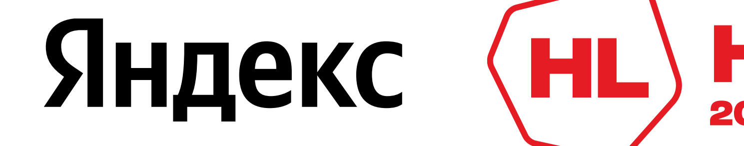
Снаружи и внутри



Кластерное ПО

- Автоматическая отработка однократных отказов
- Сохранение точки подключения
- Сохранность данных и сервиса СУБД при отказе одного узла
- Автоматическая блокировка доступа в случае отказа двух узлов для гарантированной сохранности данных






Сетевое взаимодействие



Разделение сетей:

Сети публичного сервиса

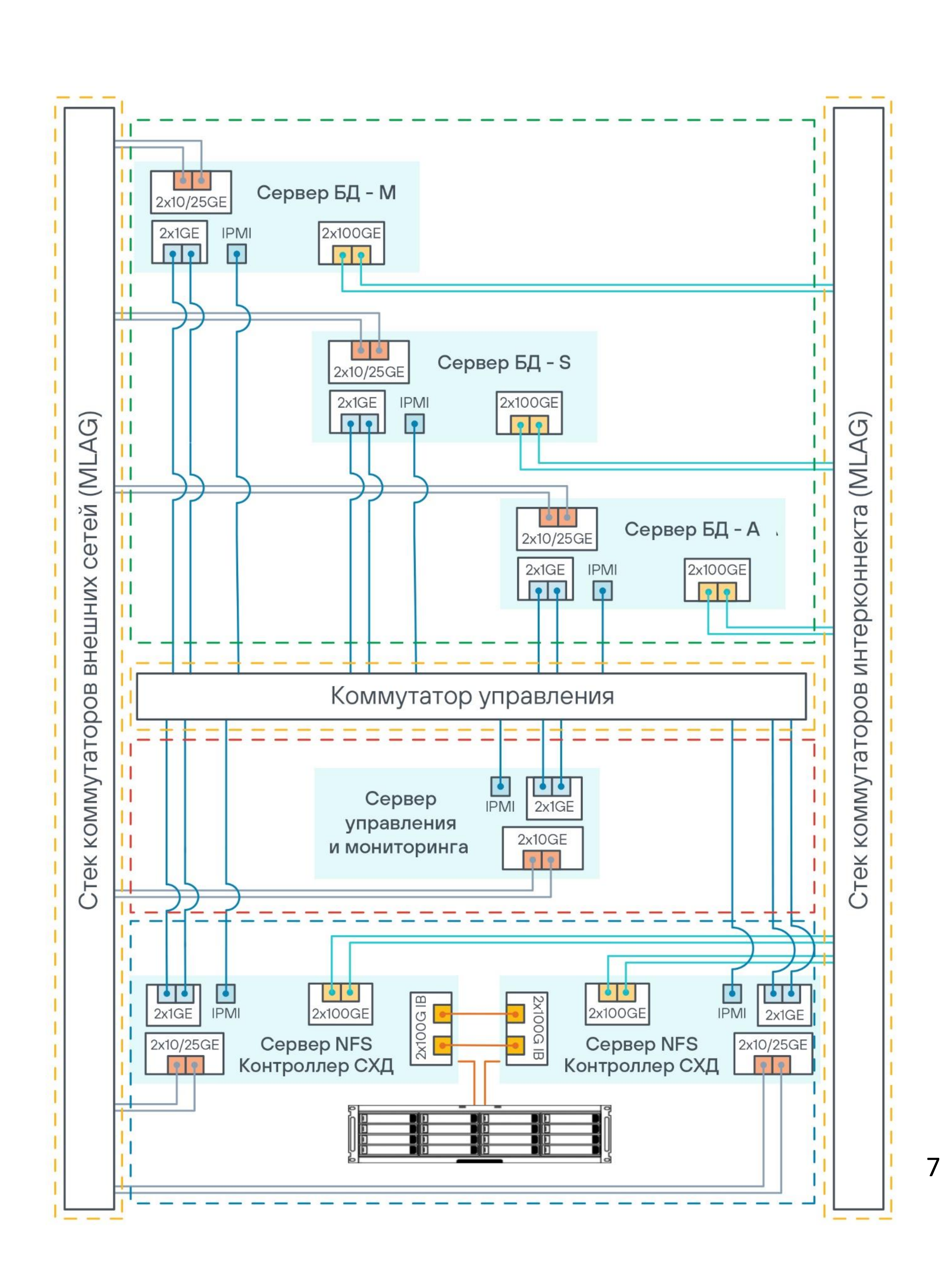
- Сеть доступа (public - default)
- Сеть внешнего бэкапа (backup)
- Сеть георепликации (dci)

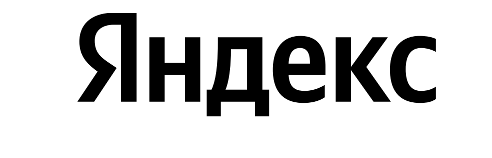
Сети управления

- Только на служебном сервере
- IPMI, MGMT (default)

Внутренние сети

- Сеть интерконнекта (internal)
- Сеть установки (PXE)
- Сеть IPMI_C (web-proxy)
- Сеть Ring (web-proxy)







Скала МБД.П вместо Oracle Exadata

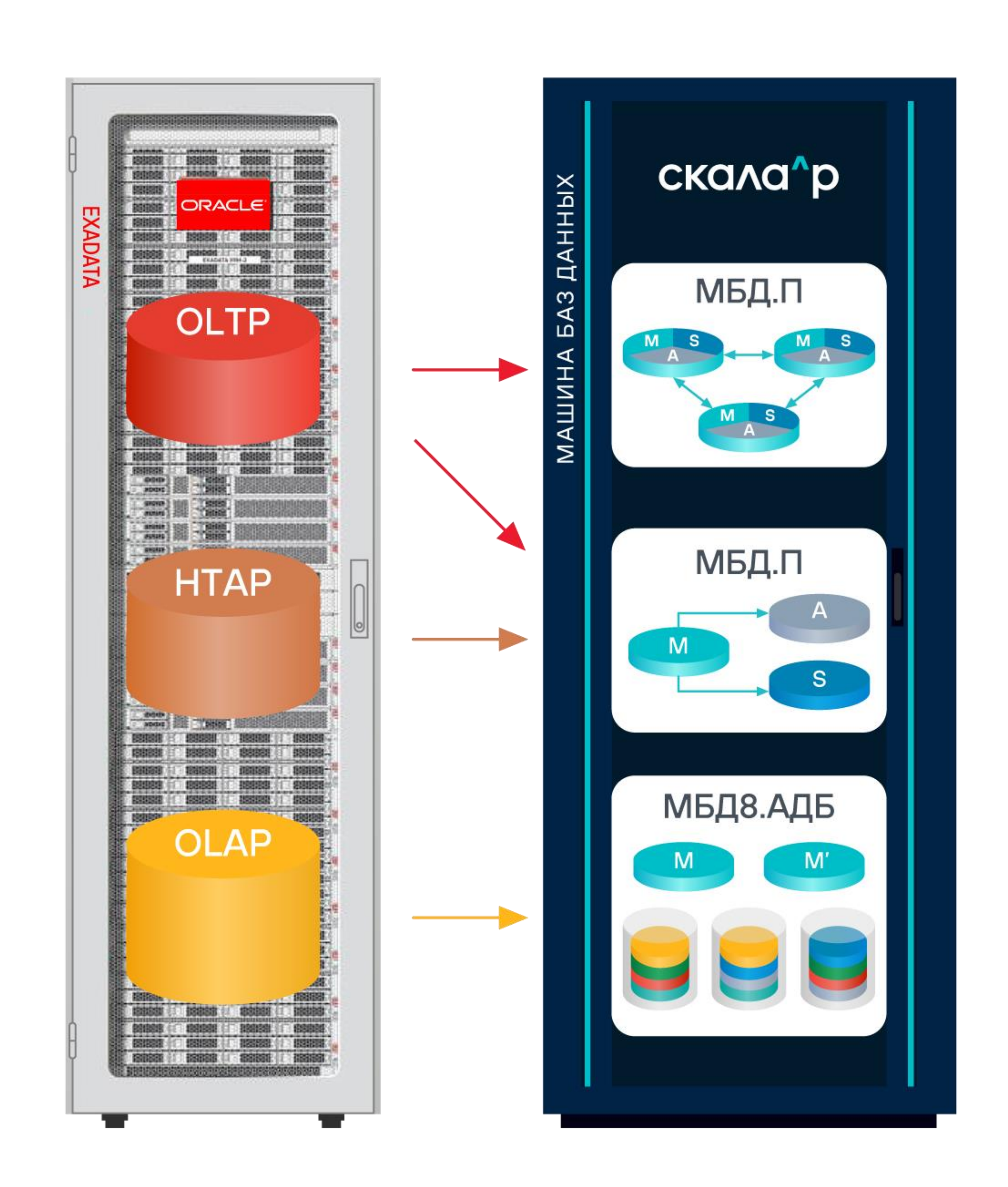


Oracle Exadata консолидирует 3 типа нагрузки:

- Транзакционная (OLTP)
- Гибридная (HTAP)
- Аналитическая (OLAP)

Транзакционная и гибридная типы нагрузок мигрируют на Скала^р МБД.П с максимально возможным уровнем производительности, доступности, сохранности данных:

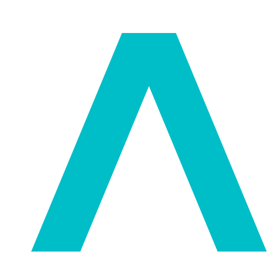
- OLTP – до 3 сервисов СУБД на кластер
- HTAP – 1 сервис СУБД, использование синхронной реплики







Типовые комплекты поставки



Модель	М-1	М-2	М-3
Параметр			
Кол-во вычислительных модулей	1	2	3
Кол-во узлов БД	3	2x3	3x3
Кол-во БД	До 3	До 6	До 9
Общий объем БД, ТБ	До 20	До 2x20	До 3x20
Объем хранения СРК, ТБ	До 130	До 260	До 260



Скорость дисковой подсистемы



Программный рейд

- Производительнее аппаратного RAID-контроллера
- Минимальное использование RAM (требуется менее 4GB RAM)
- Управление процессорными потоками
- Минимальная просадка производительности в режиме восстановления
- Мониторинг износа накопителя
- Автозамена накопителя
- Повышение уровня RAID

fio -ioengine=libaio -direct=1 -bs=4k -iodepth=1 -fsync=1 -rw=randwrite

ERA Raidix RAID10:
- write: IOPS=20.7k, BW=80.9MiB/s (84.8MB/s);
- lat (usec): min=34, max=756, avg=44.6

mdadm RAID10:
- write: IOPS=16.8k, BW=65.5MiB/s (68.7MB/s);
- lat (usec): min=35, max=671, avg=55.52

LSI MegaRAID RAID10:
- write: IOPS=19.1k, BW=73.9MiB/s (77.4MB/s);
- lat (usec): min=41, max=528, avg=48.27





Сети. Использование RDMA



USER
KERNEL
HARDWARE

Приложение 1
Buffer 1
ОС
Buffer 1
Buffer 1
NIC
Buffer 1
RACK 1

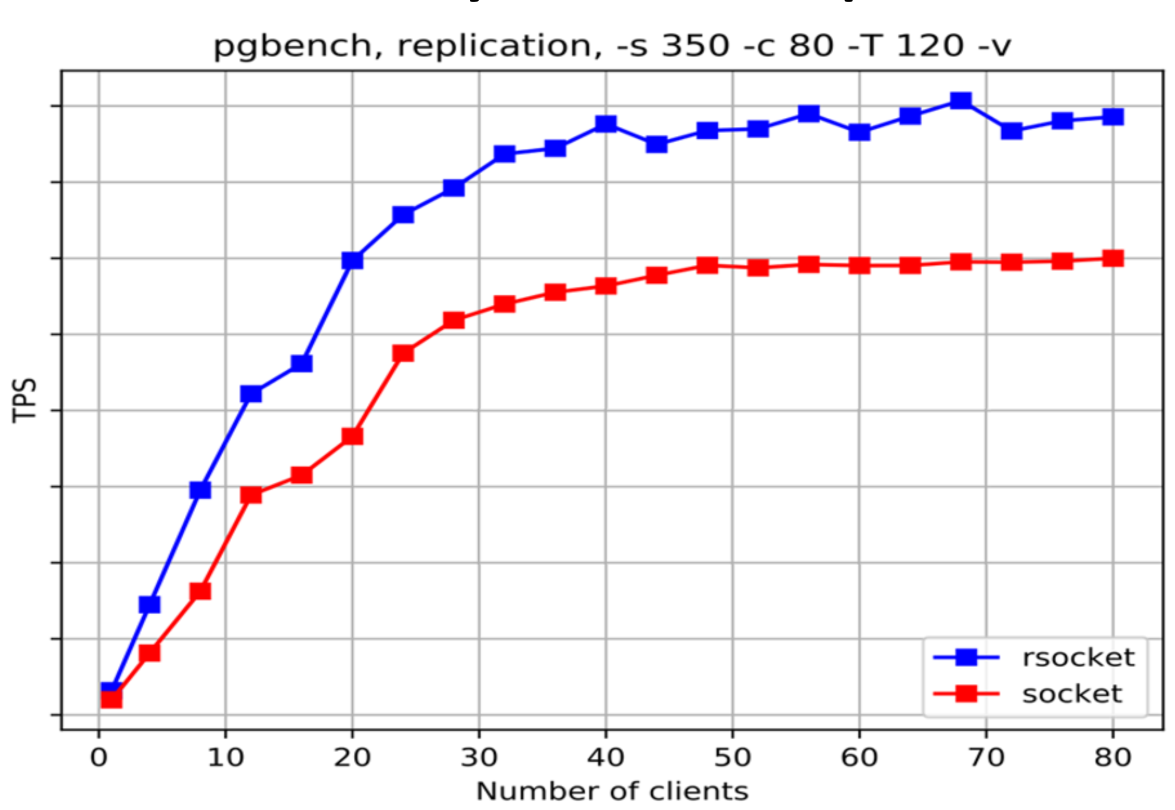
RDMA поверх InfiniBand или Ethernet
NIC
NIC
TCP/IP

Приложение
Buffer 1
Buffer 1
ОС
Buffer 1
Buffer 1
NIC
RACK 2

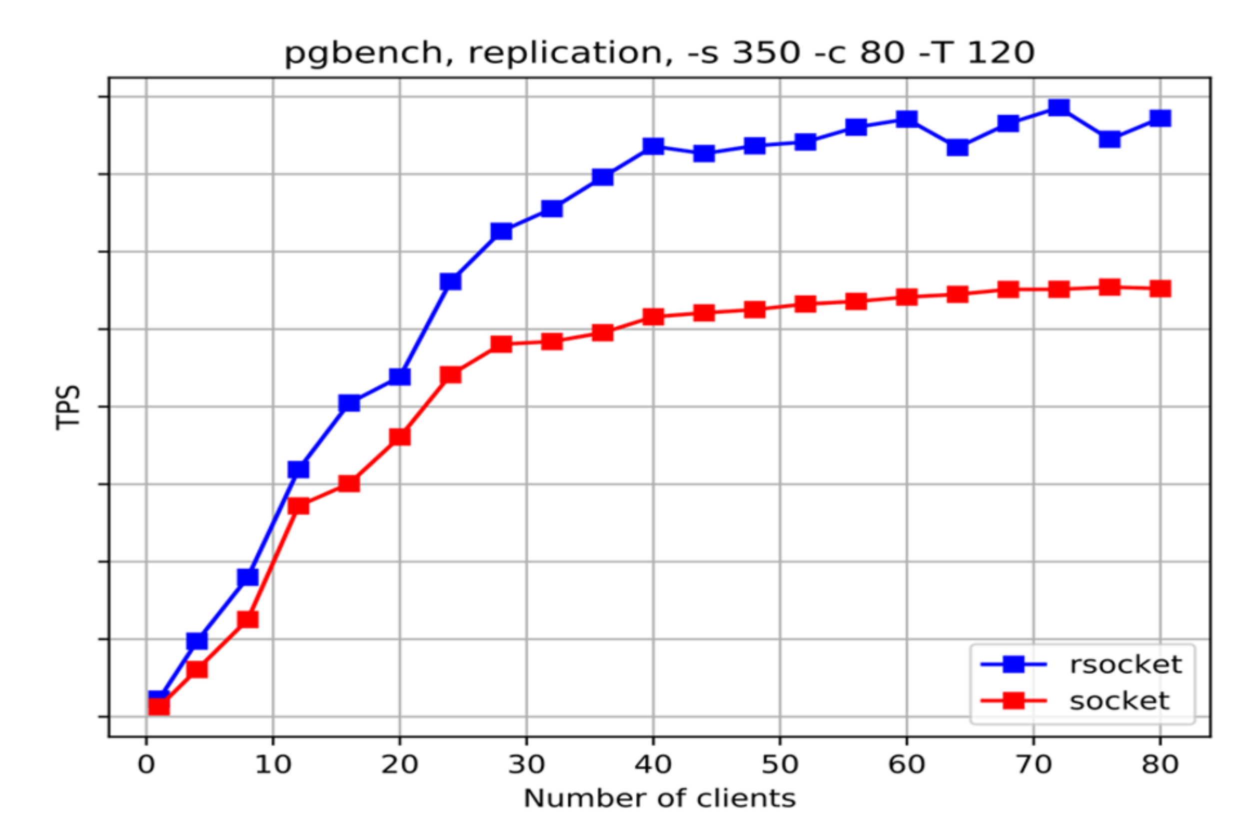
Сети. Использование RDMA



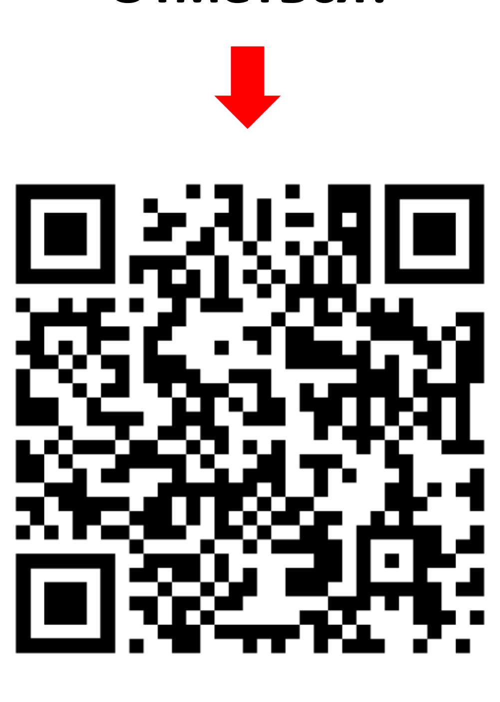
Read-write with synchronous replication, TPS



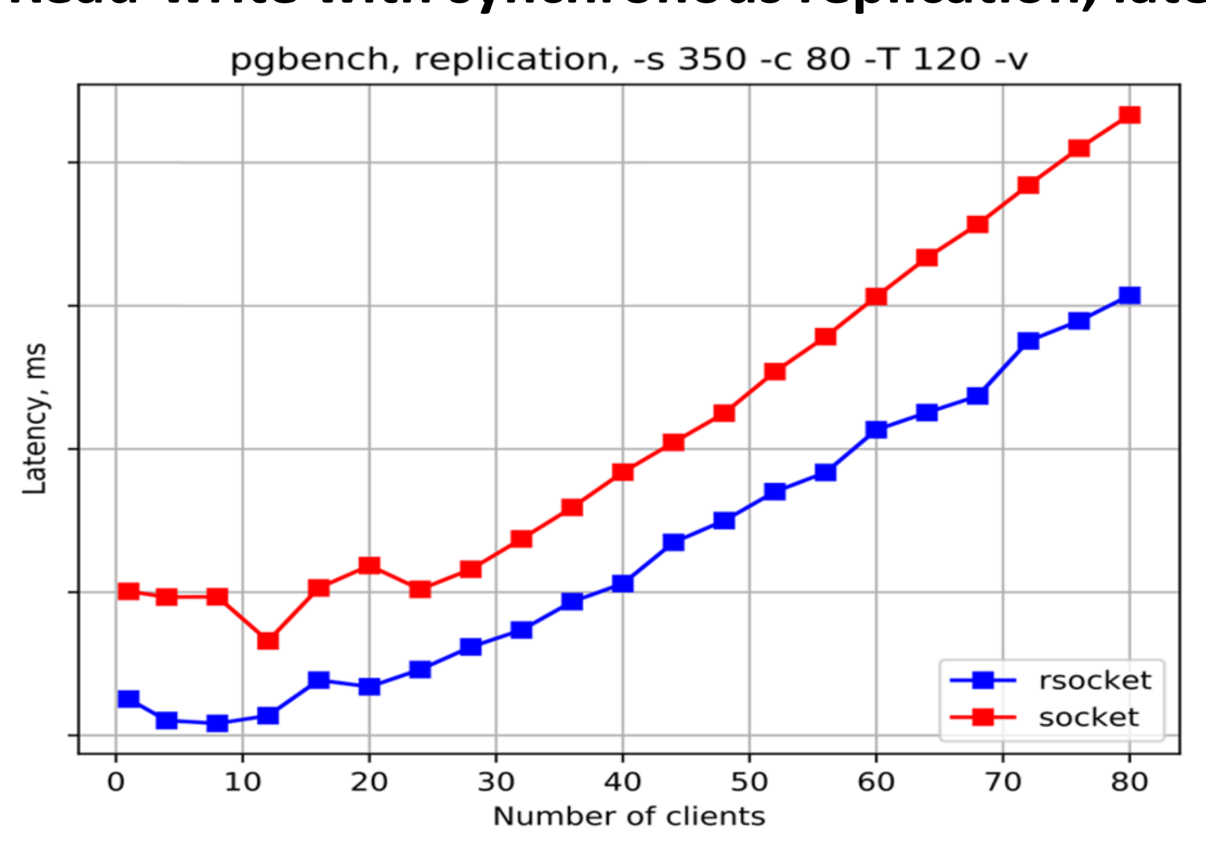
Read-write with remote apply replication, TPS



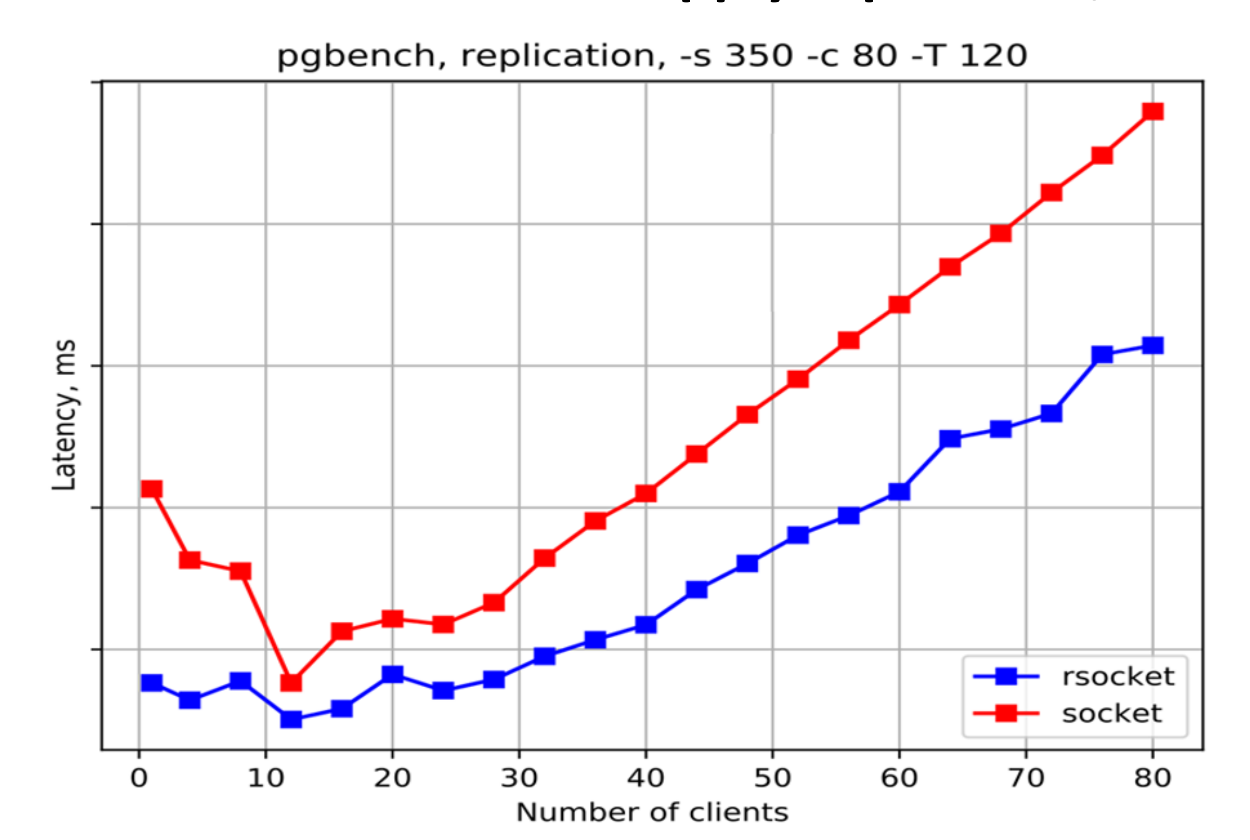
Знаешь Скала^р?
Отметься!



Read-write with synchronous replication, latency



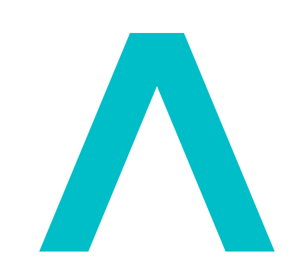
Read-write with remote apply replication, latency







Сети. Не все коммутаторы одинаково полезны



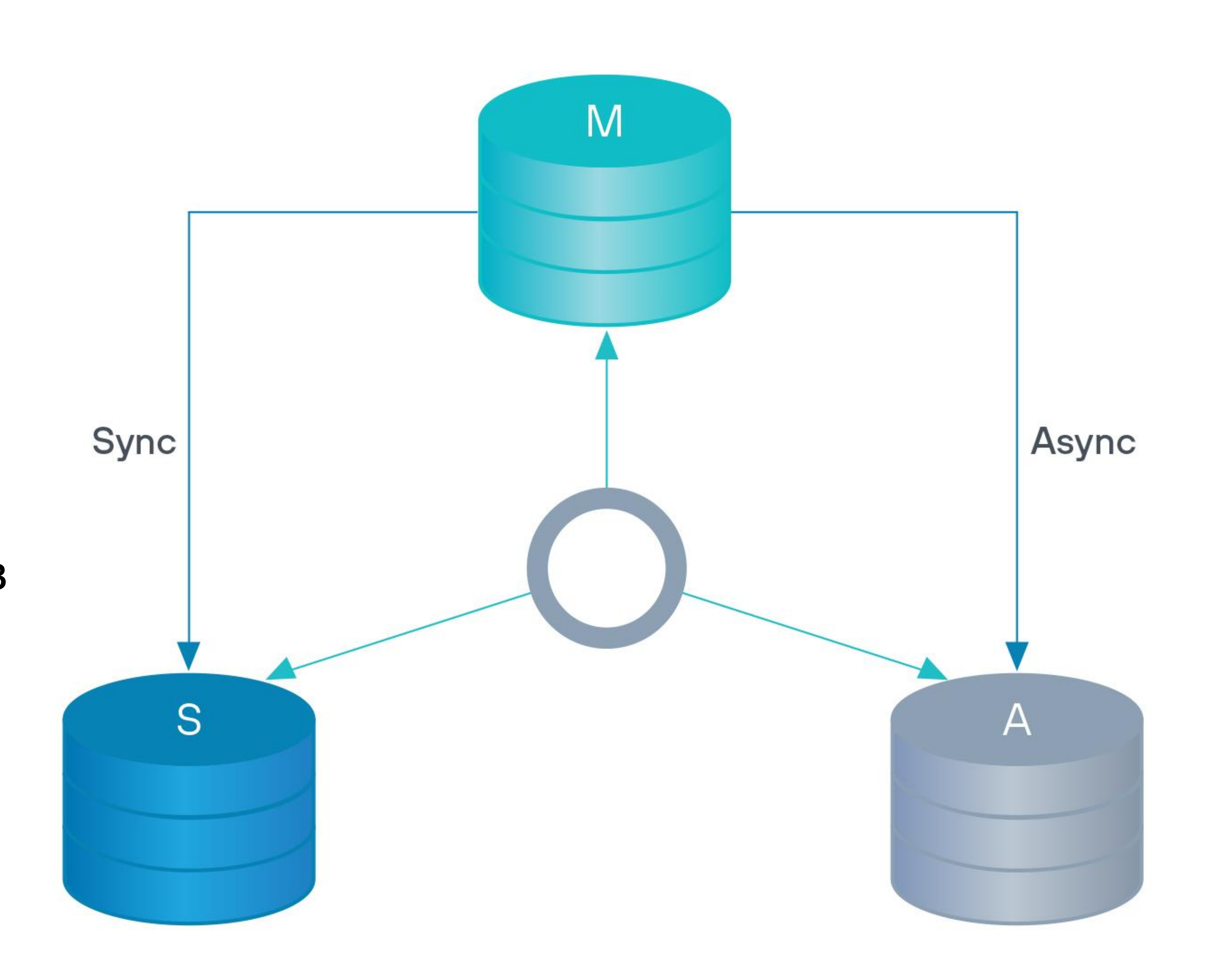
Производитель	Модель	Время сходимости	Пропускная способность	VLAN	RSTP	RDMA	Порты	Производство	Цена
Vendor#1	Model#1	**100**	92	4000	On	+/-	32	РФ	0.8X
Vendor#2	Model#2	**2000**	96	4096	Off	-	32	РФ	2.7X
Vendor#3	Model#3	**3000**	100	4096	?	+	32	Китай	2.5X
Vendor#4	Model#4	250	100	4096	On	+	32	Китай	1.5X
Mellanox	SN2100	40	100	4096	On	+	16	США	X
Vendor#5	Model#5	-	100	4000	On	-	32	РФ	???

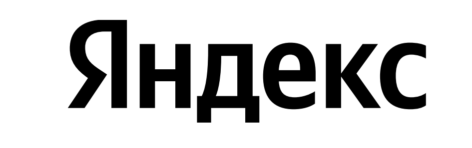
Отказоустойчивость



Отказоустойчивость на всех уровнях

- Надежные комплектующие
- Резервирование значимых компонентов на аппаратном уровне
- Отказоустойчивая архитектура верхнего уровня
  - Оперативное восстановление при сбоях
  - Автоматическая отработка однократных отказов
- Сохранение точки подключения
- Сохранность данных и сервиса СУБД при отказе одного узла
- Автоматическая блокировка доступа в случае отказа двух узлов для гарантированной сохранности данных





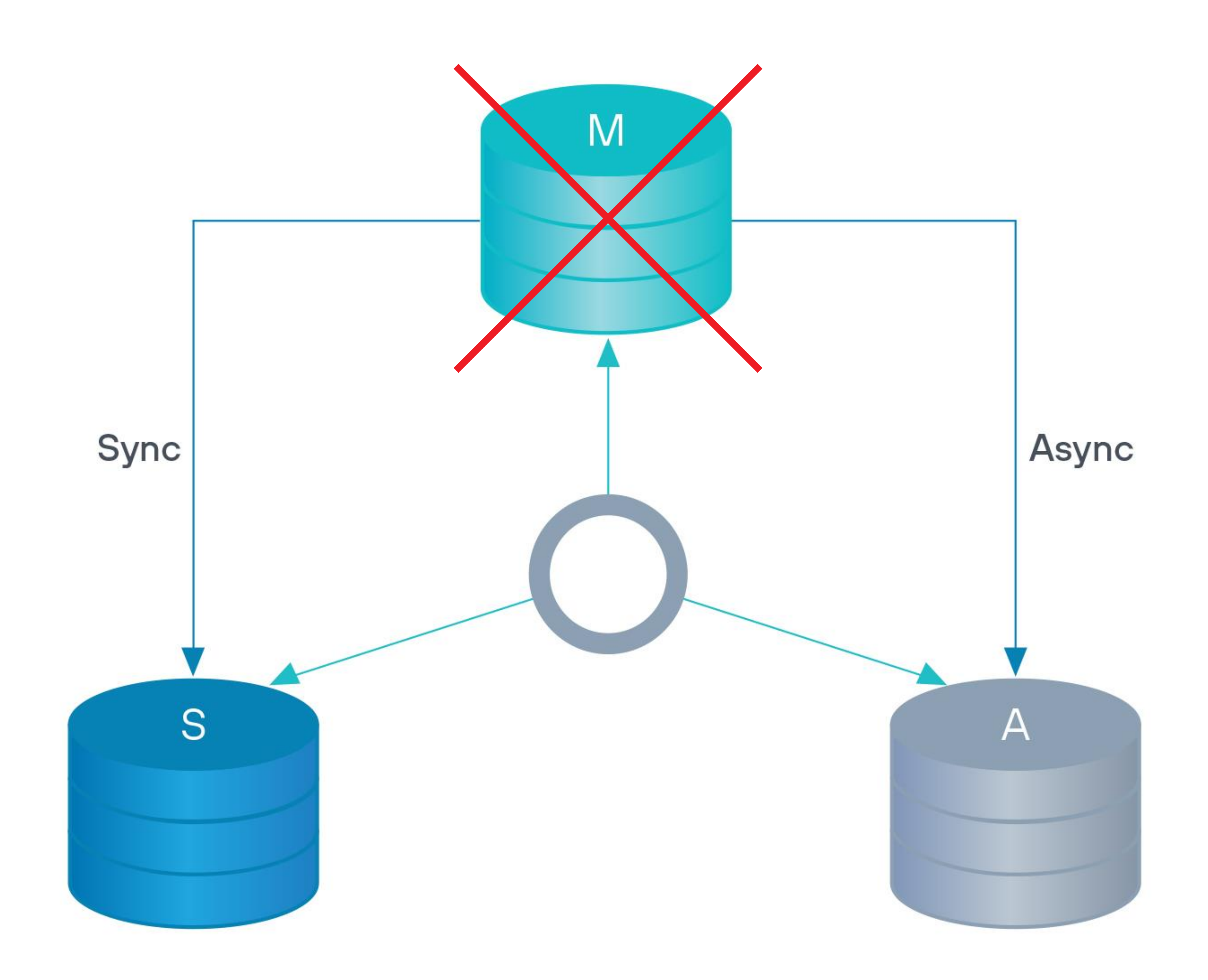


Отказоустойчивость



Отказоустойчивость на всех уровнях

- Отказоустойчивая архитектура верхнего уровня
  - Оперативное восстановление при сбоях
  - Автоматическая отработка однократных отказов
- Сохранение точки подключения





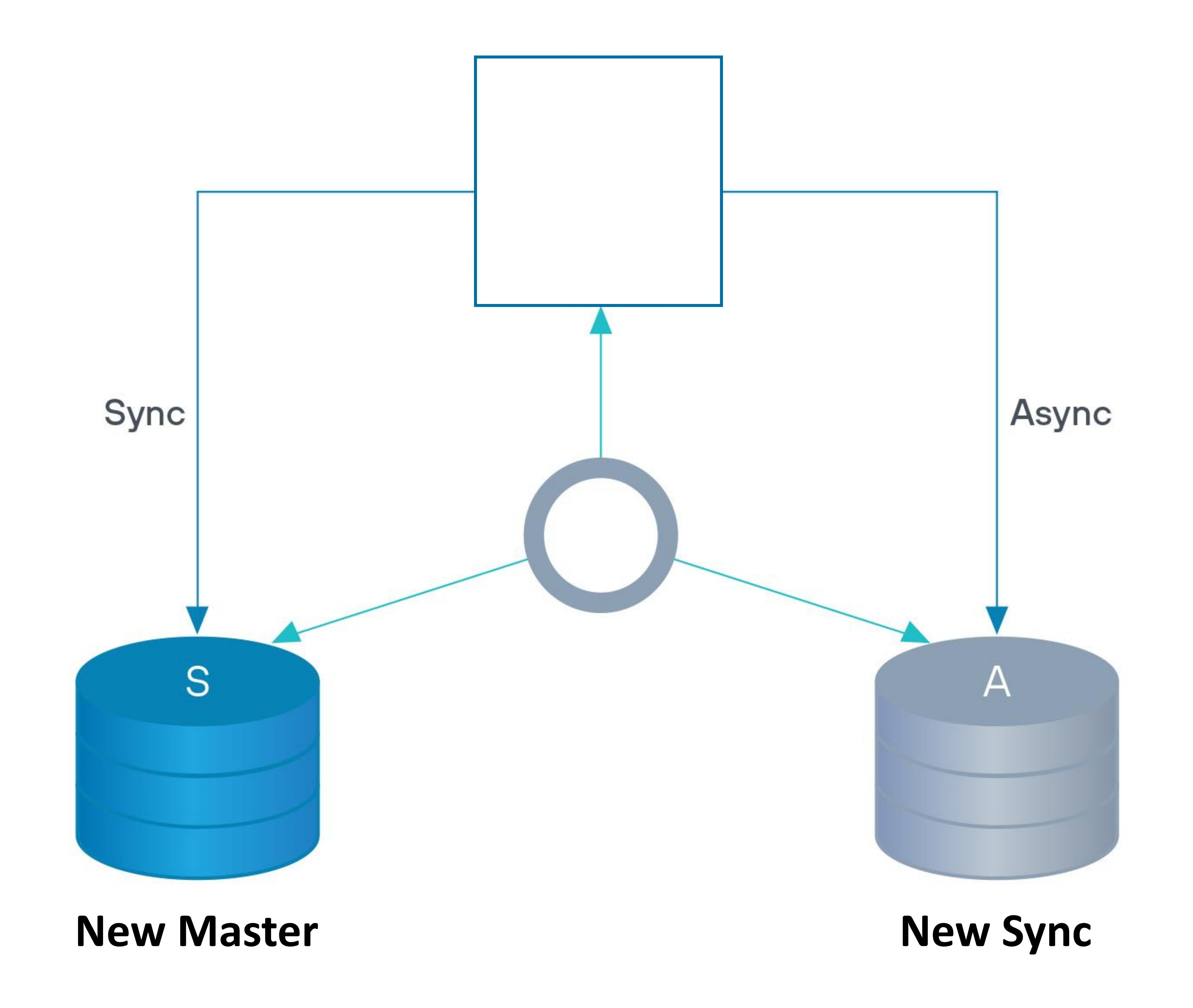


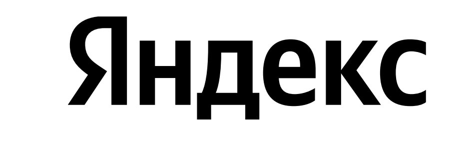
Отказоустойчивость



Отказоустойчивость на всех уровнях

- Сохранение точки подключения
- Сохранность данных и сервиса СУБД при отказе одного узла
- Автоматическая блокировка доступа в случае отказа двух узлов для гарантированной сохранности данных

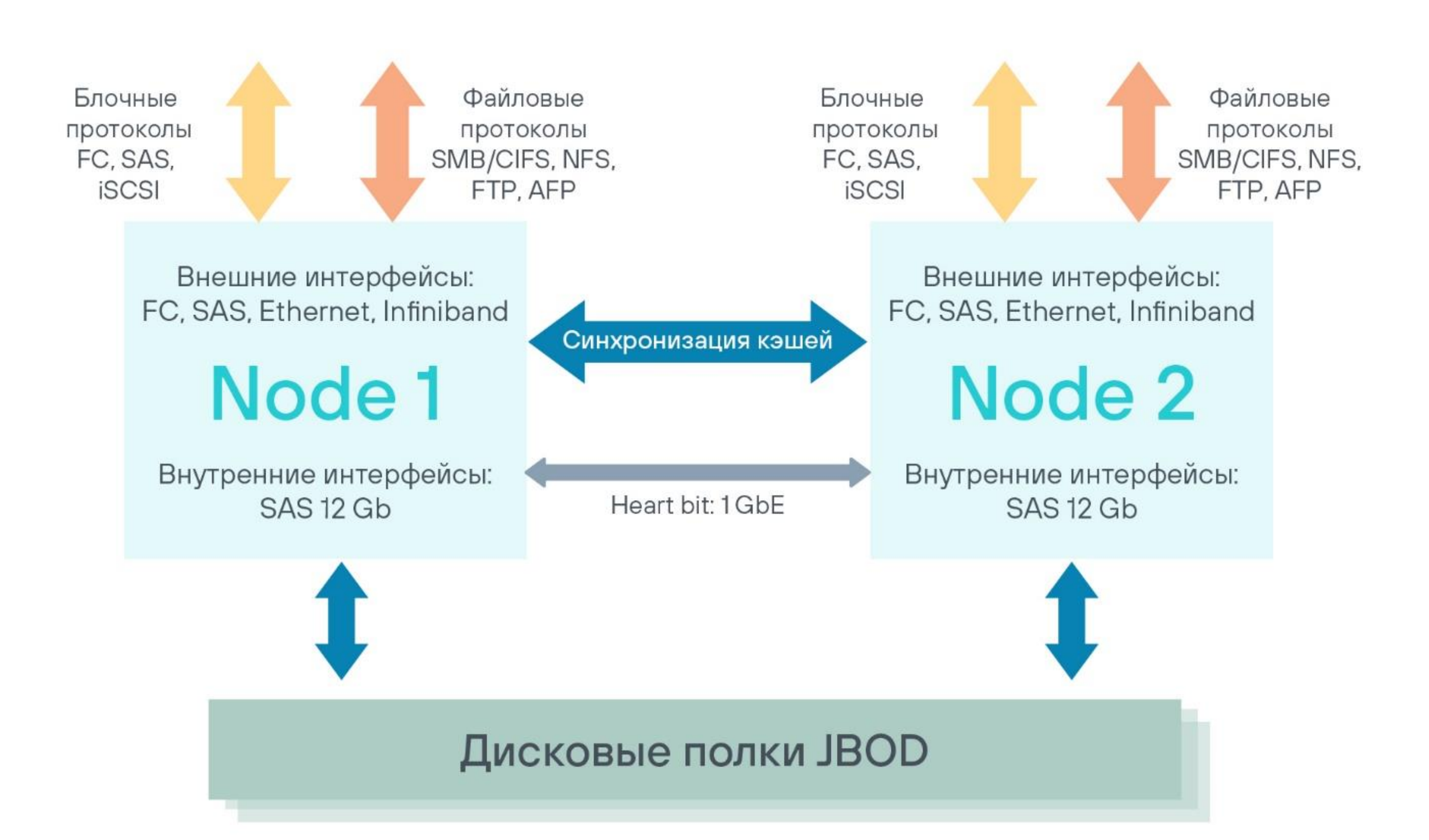


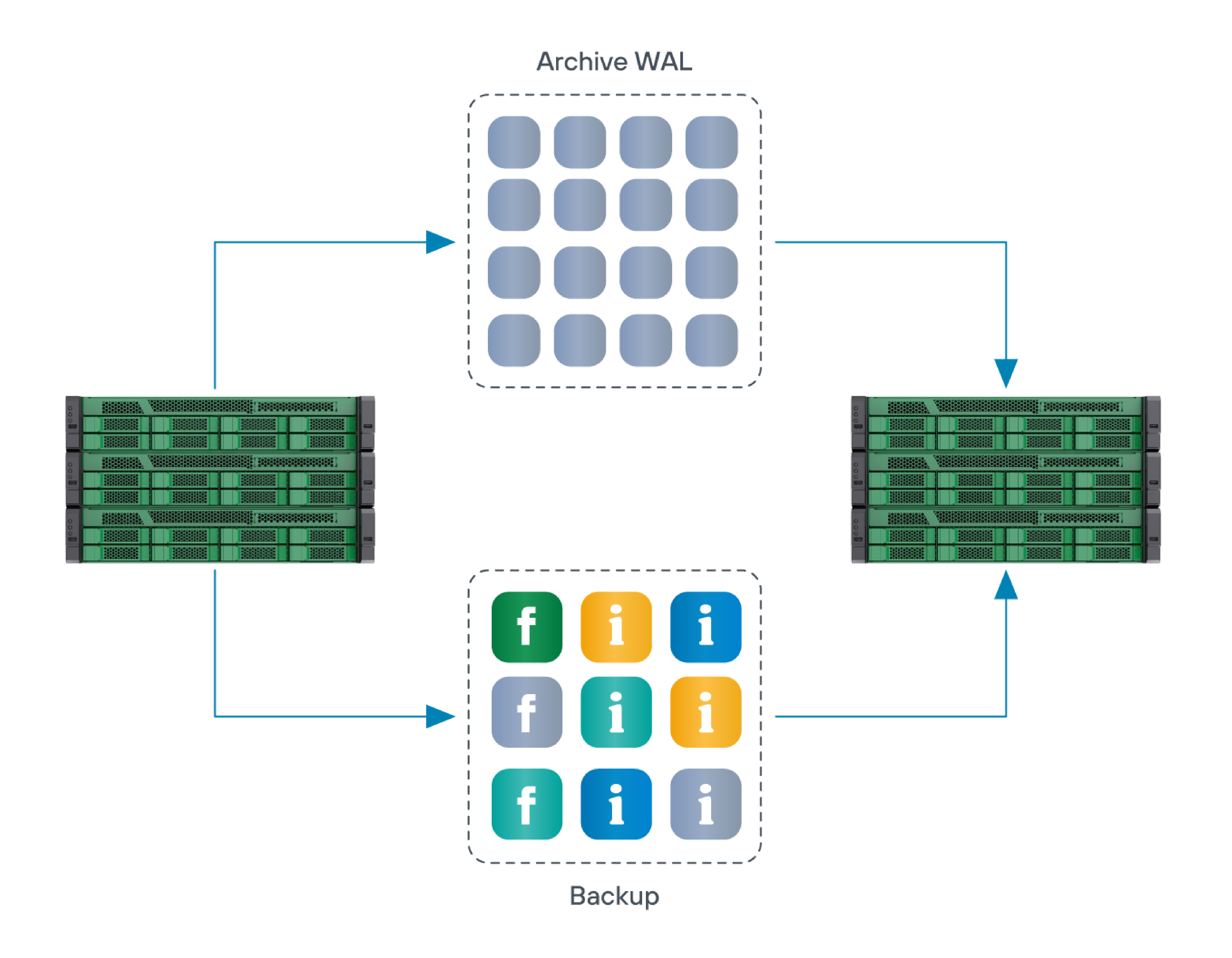




Резервное копирование











Меряем производительность

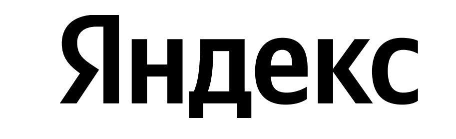


Измеряем правильно!
Андрей Николаенко, PG Day’17 –
«Об эталонном тестировании PostgreSQL»



- pgbench, TPC-B-подобная нагрузка
- SF=70000
- RW-тест с БД объемом более 1 Тб
- SO-тест с тома объемом 12 Тб

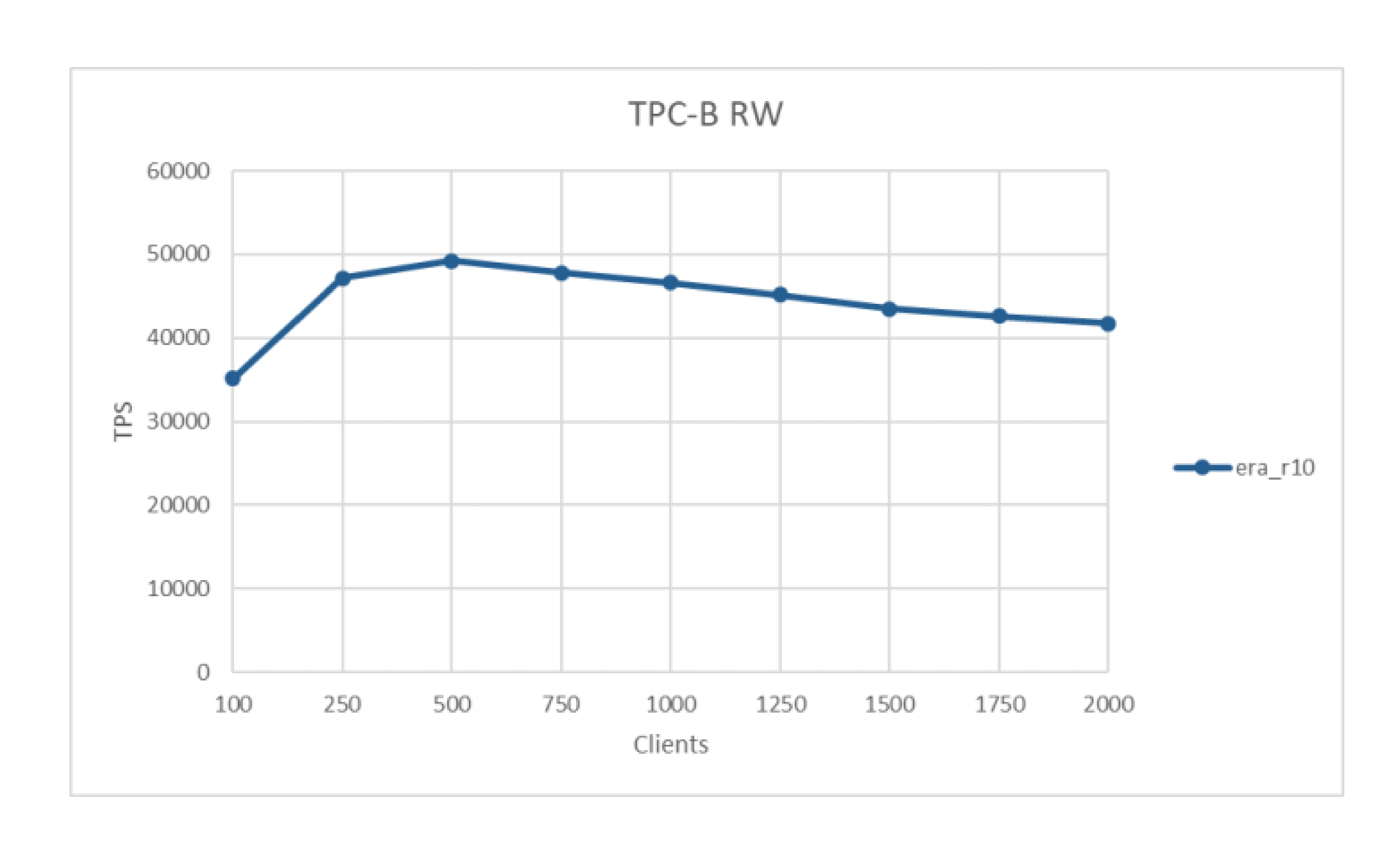
```
1.BEGIN;
2.UPDATE pgbench_accounts SET abalance = abalance + :delta WHERE aid = :aid;
3.SELECT abalance FROM pgbench_accounts WHERE aid = :aid;
4.UPDATE pgbench_tellers SET tbalance = tbalance + :delta WHERE tid = :tid;
5.UPDATE pgbench_branches SET bbalance = bbalance + :delta WHERE bid = :bid;
6.INSERT INTO pgbench_history (tid, bid, aid, delta, mtime) VALUES (:tid, :bid, :aid, :delta, CURRENT_TIMESTAMP);
7.END;
```

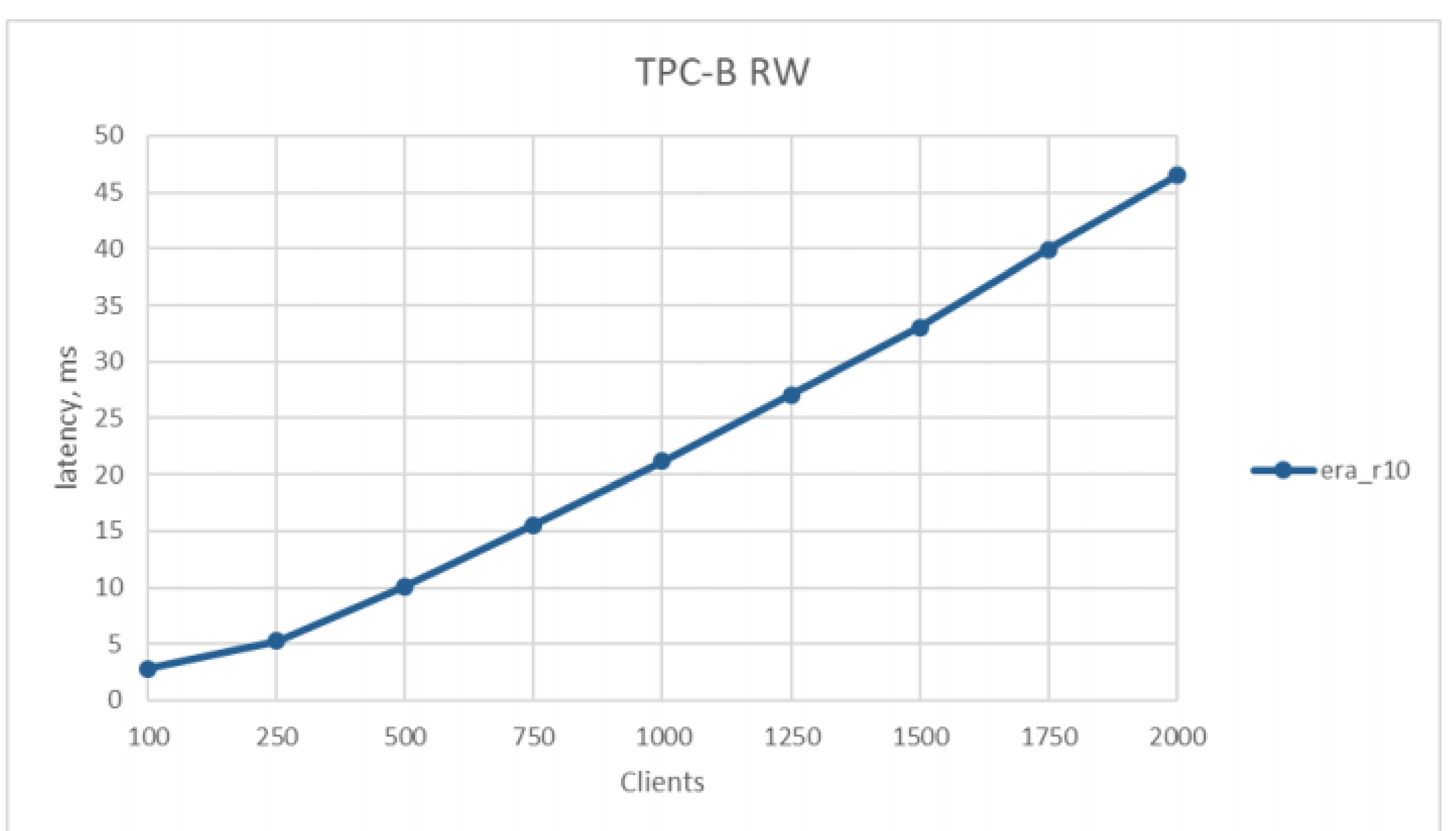





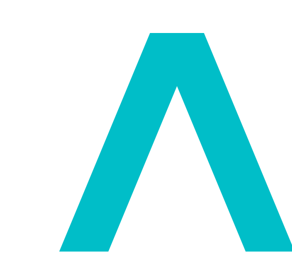
Меряем производительность

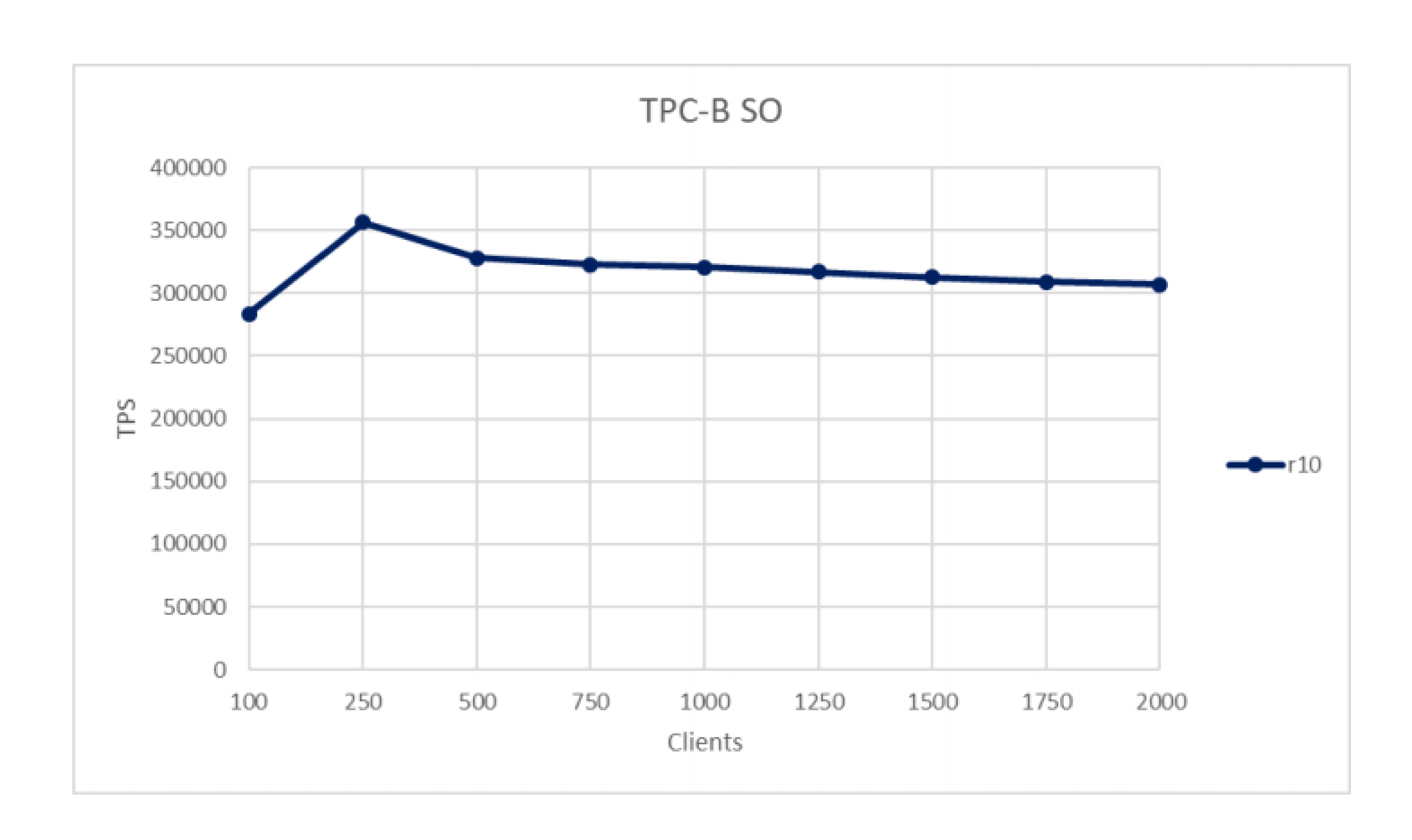


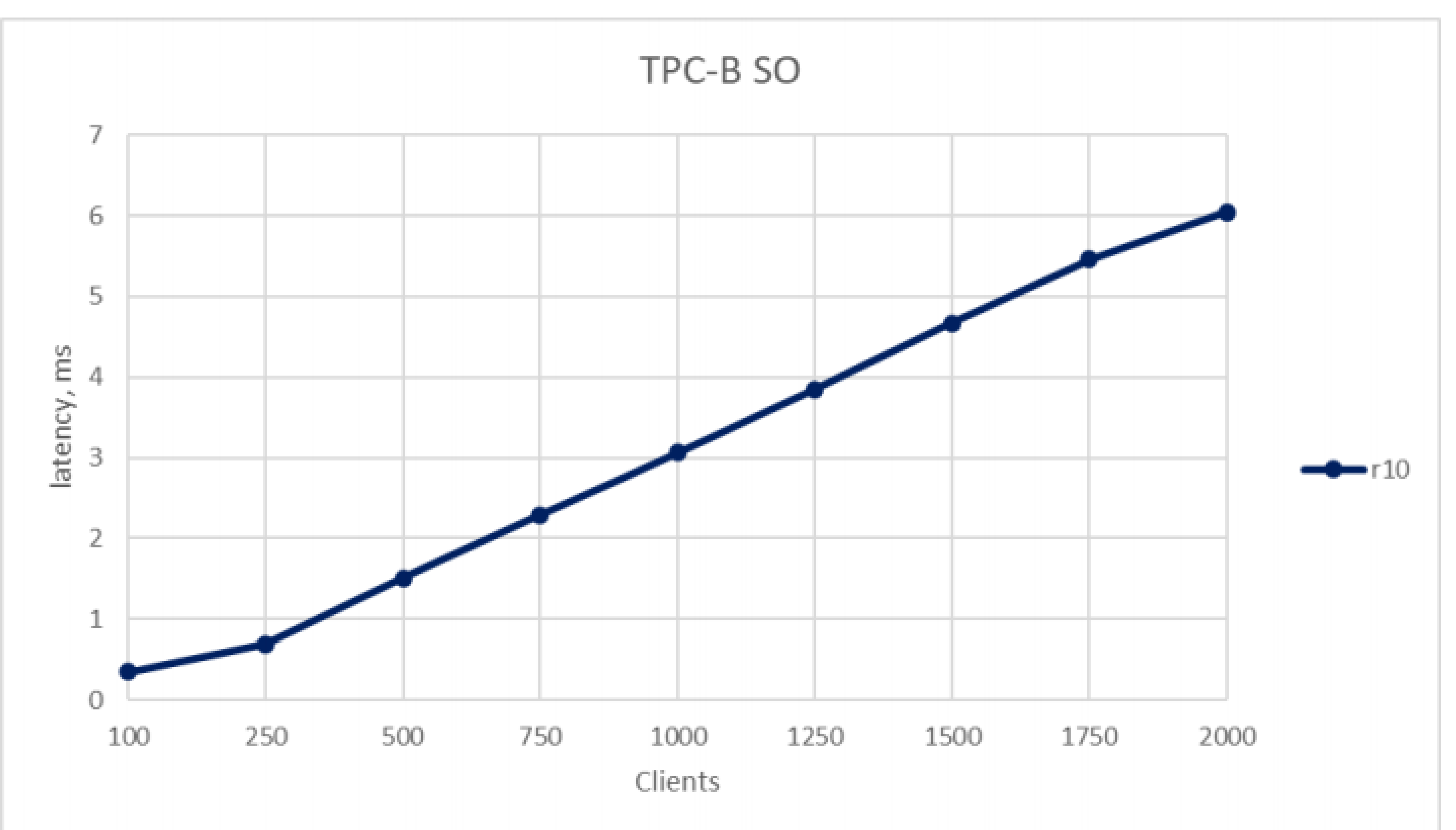




Меряем производительность



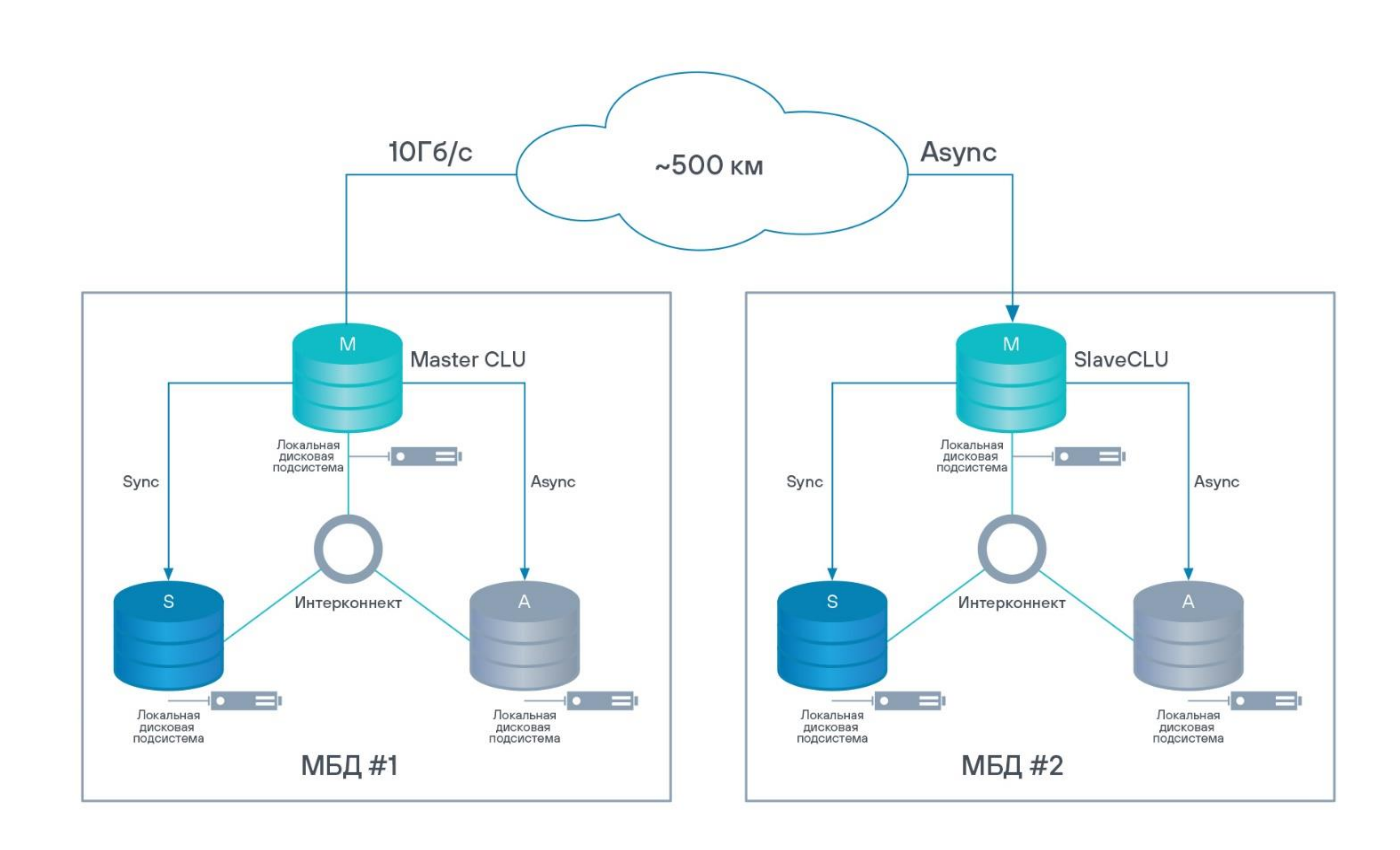




Случай из жизни №1. Распределенный геокластер



- Вторая, взаимодействующая «Скала^р МБД #2» (полная копия основной)
- Удаленность порядка 500 км
- Канал 10 Гбит/с
- Асинхронная реплика основного кластера
- Каскадная репликация

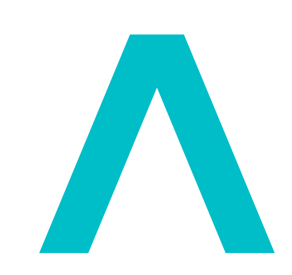


- Вечерний пятничный полный бэкап ~11 ТБ (сжатие) за 16,5 часов (режим ARCHIVE) + 500 ГБ архивных журналов
- Утренний инкрементальный бэкап в понедельник ~200–400 ГБ за 30–50 минут (режим PAGE) + 10–20 ГБ архивных журналов
- Инкрементальный бэкап каждые 2 часа в рабочие дни ~10–30 ГБ за ~3–5 минут (режим PAGE) + 5–15 ГБ архивных журналов



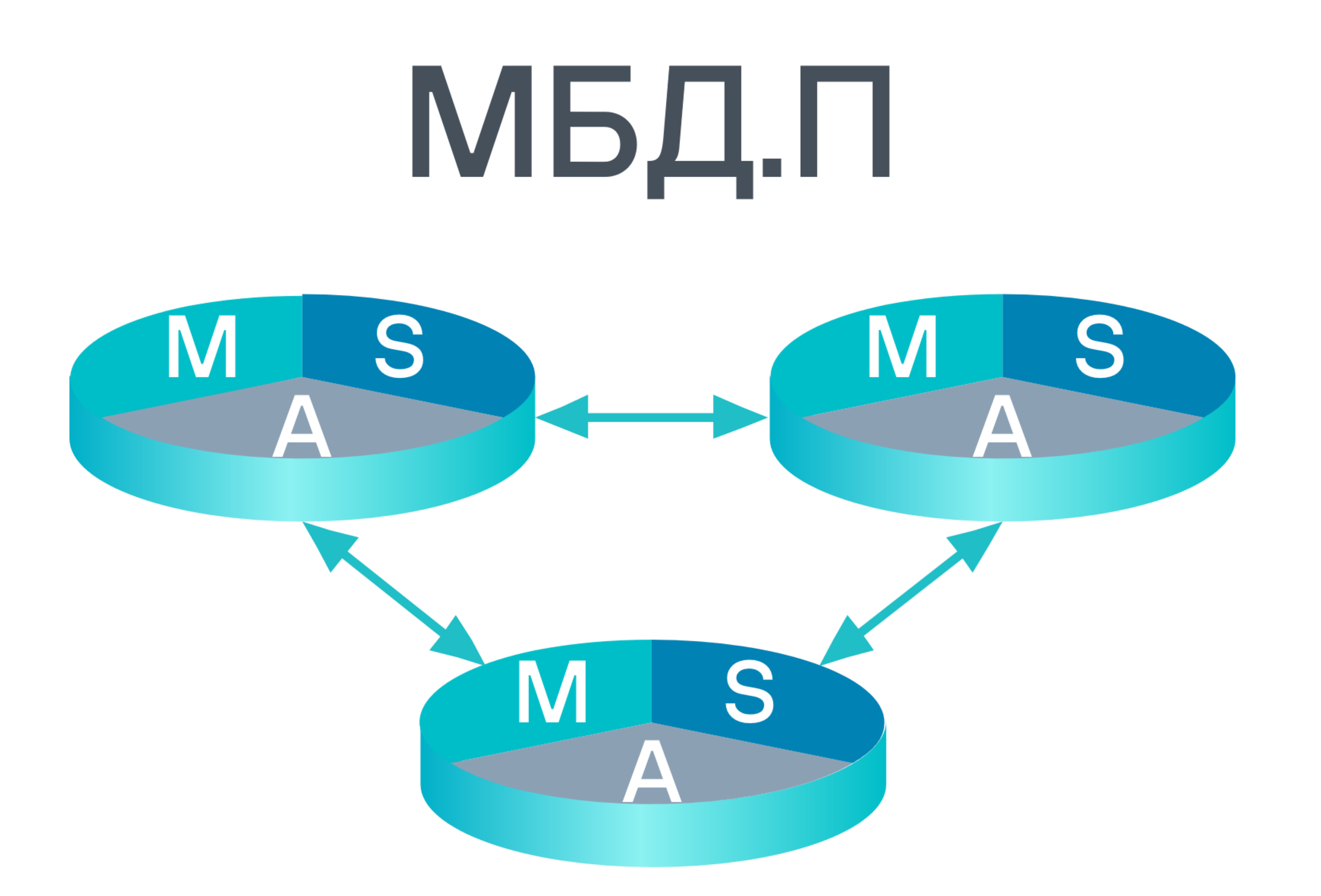


Случай из жизни №2. «Шахматка»



Решение – Скала^р МБД.П:

- 1 аппаратный кластер (3 узла)
- 3 экземпляра БД
- СУБД Postgres Pro Enterprise
- Объем БД № 1 ~18 ТБ
- Объем БД № 2 ~17,5 ТБ
- Объем БД № 3 ~15 ТБ

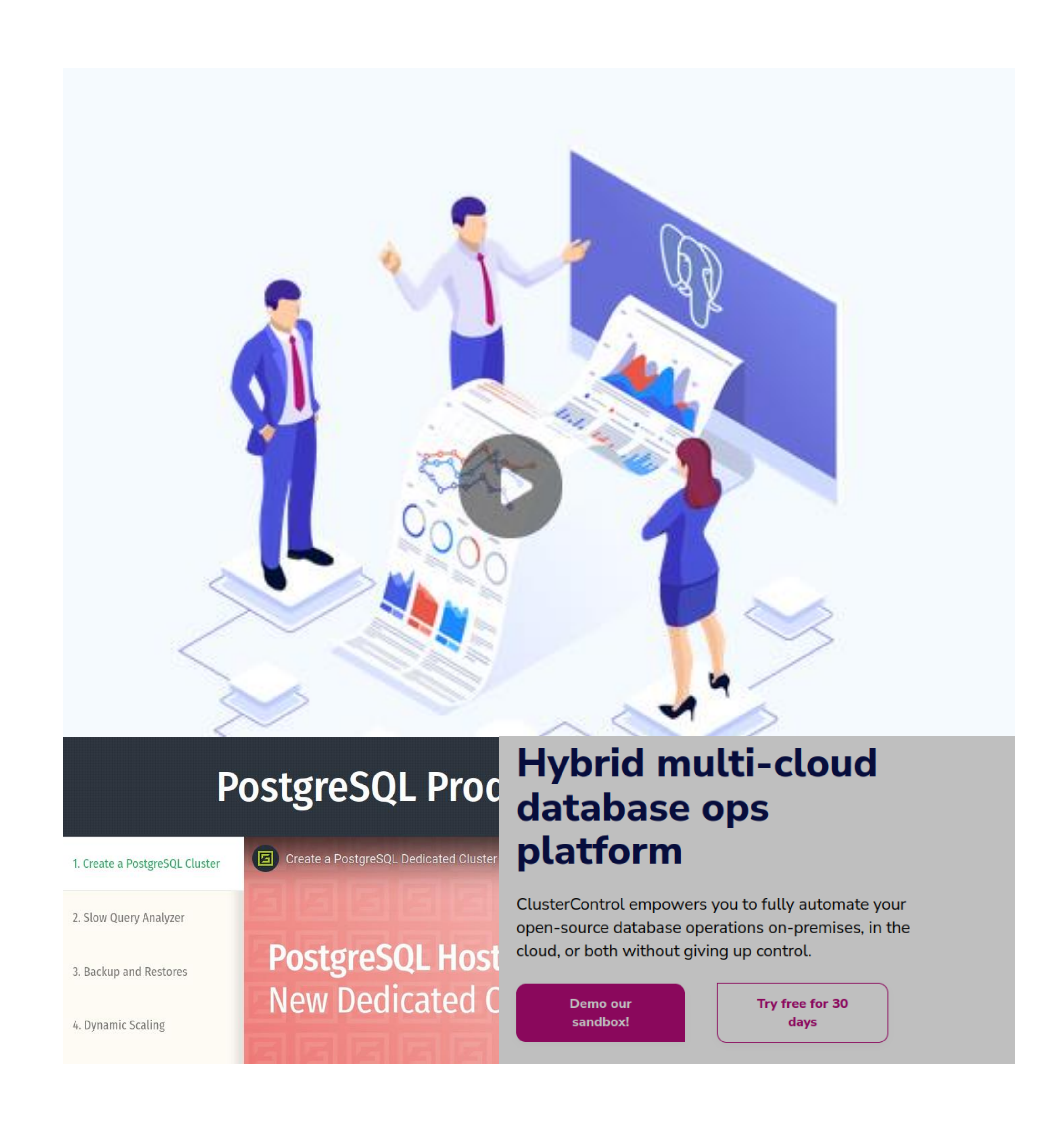


Система управления Машиной БД



Что есть:

- **SeveralNines Cluster Control**
  https://severalnines.com/product/clustercontrol
- **ScaleGrid**
  https://scalegrid.io/postgresql/demo.html
- **Awide**
  https://awide.io/





Система управления Машиной — ^Спектр!



^спектр

Управление

- Кластерами и узлами МДБ.П
- Отказоустойчивыми сервисами СУБД
- Резервным копированием и восстановлением БД (в разработке)

Мониторинг

- Серверного оборудования
- Сервисов СУБД
- Сетевого оборудования (в разработке)
- Резервного копирования (в разработке)

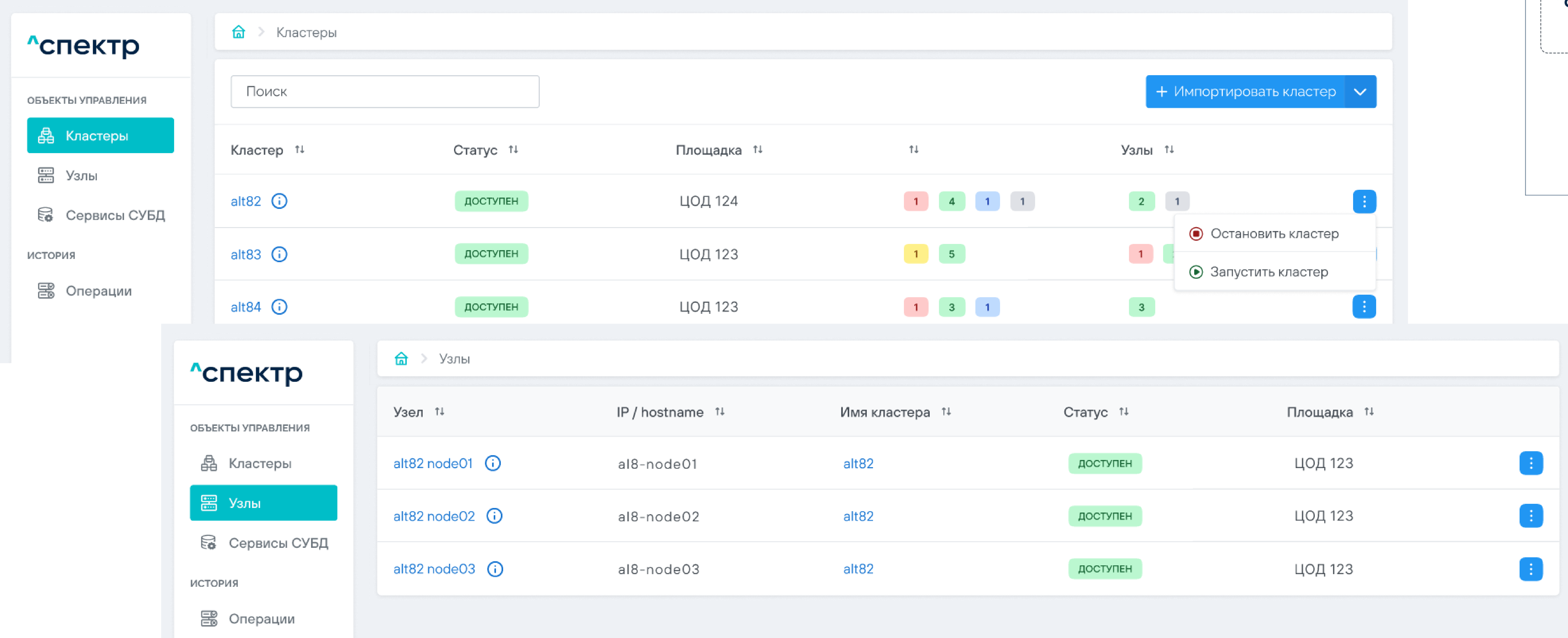


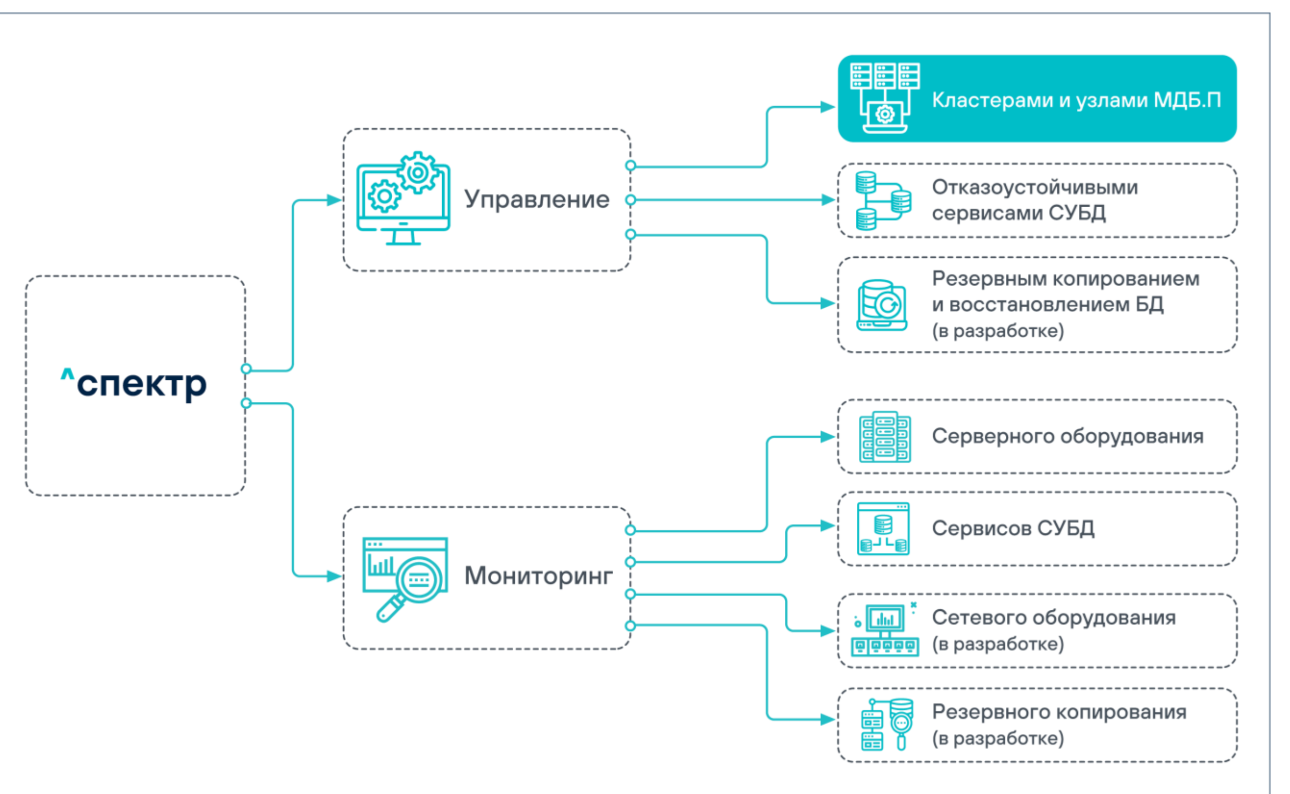


Система управления Машиной — ^Спектр!



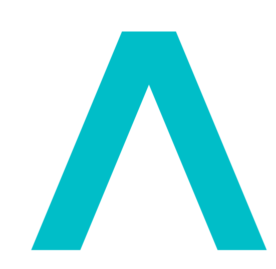
Снижает сложность операций с кластерным ПО
Использует проверенные шаблоны настроек



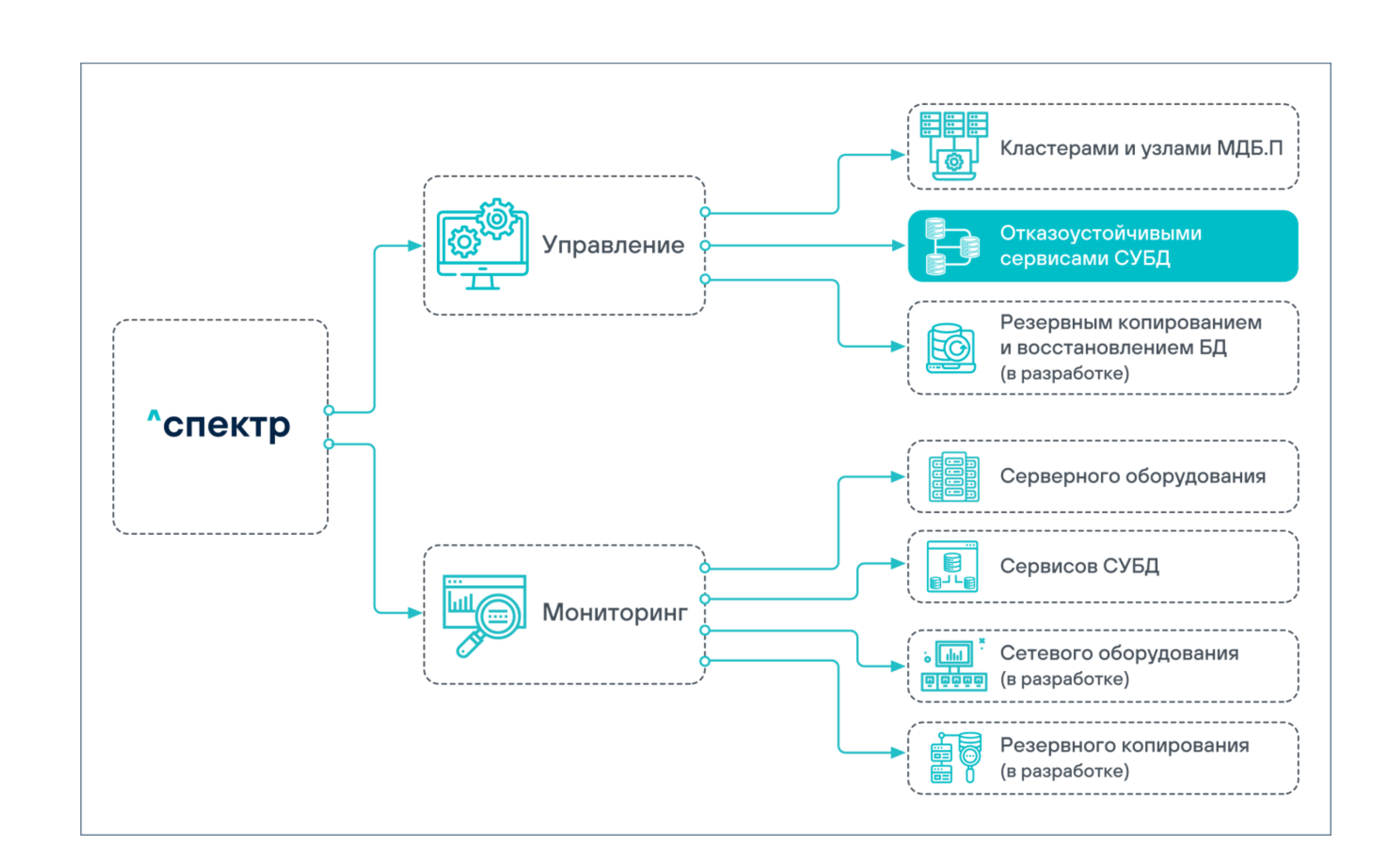


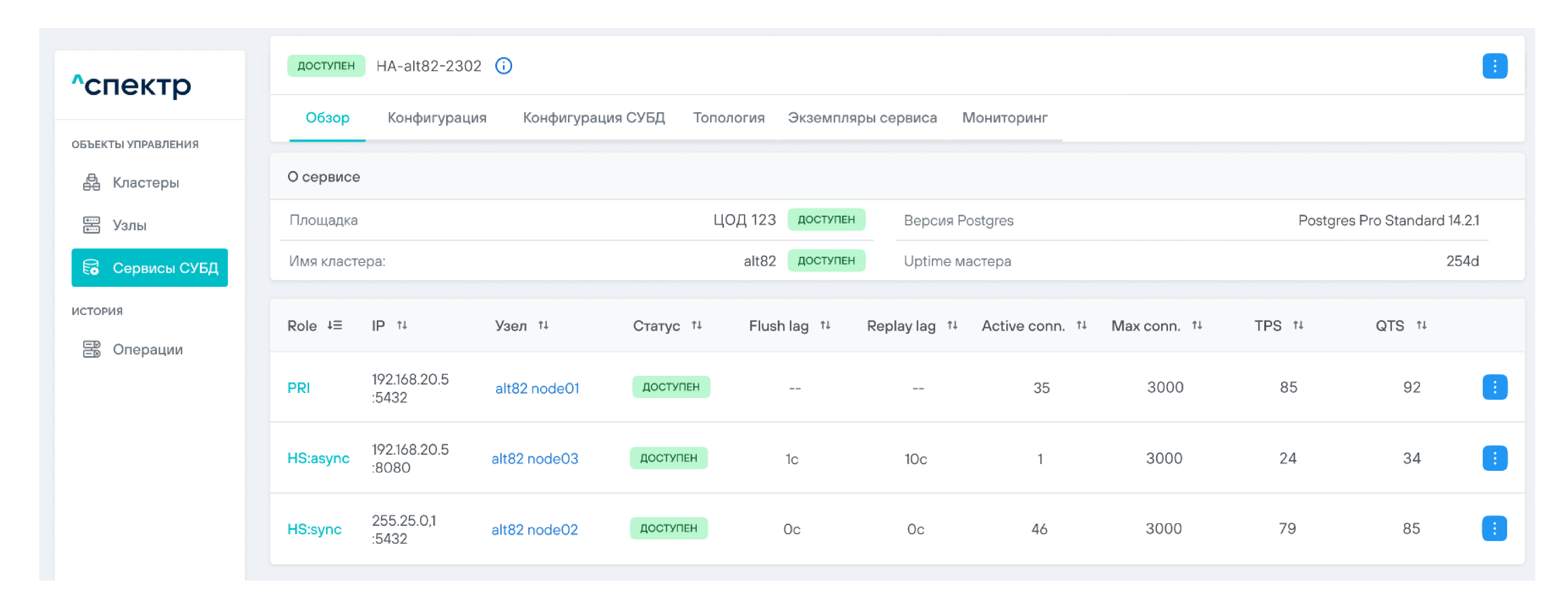


Управление и настройка сервисов в СУБД



Показывает статус и взаимосвязи сервисов СУБД
Акцентирует внимание на неисправных объектах
Автоматизирует операцию switchover сервиса СУБД









Мониторинг серверного оборудования



^спектр

ОБЪЕКТЫ УПРАВЛЕНИЯ
Кластеры
Узлы
Сервисы СУБД

ИСТОРИЯ
Операции

Настройки

Узлы > alt82 node01

ДОСТУПЕН ЦОД 123 / alt82 node01

Обзор | Экземпляры сервиса | Мониторинг

1h 3h 1d 1w 1mo 2021-12-06 16:45 — 2021-12-07 16:45 Мониторинг МБД

18:00 20:00 22:00 2021-12-07 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 Сейчас

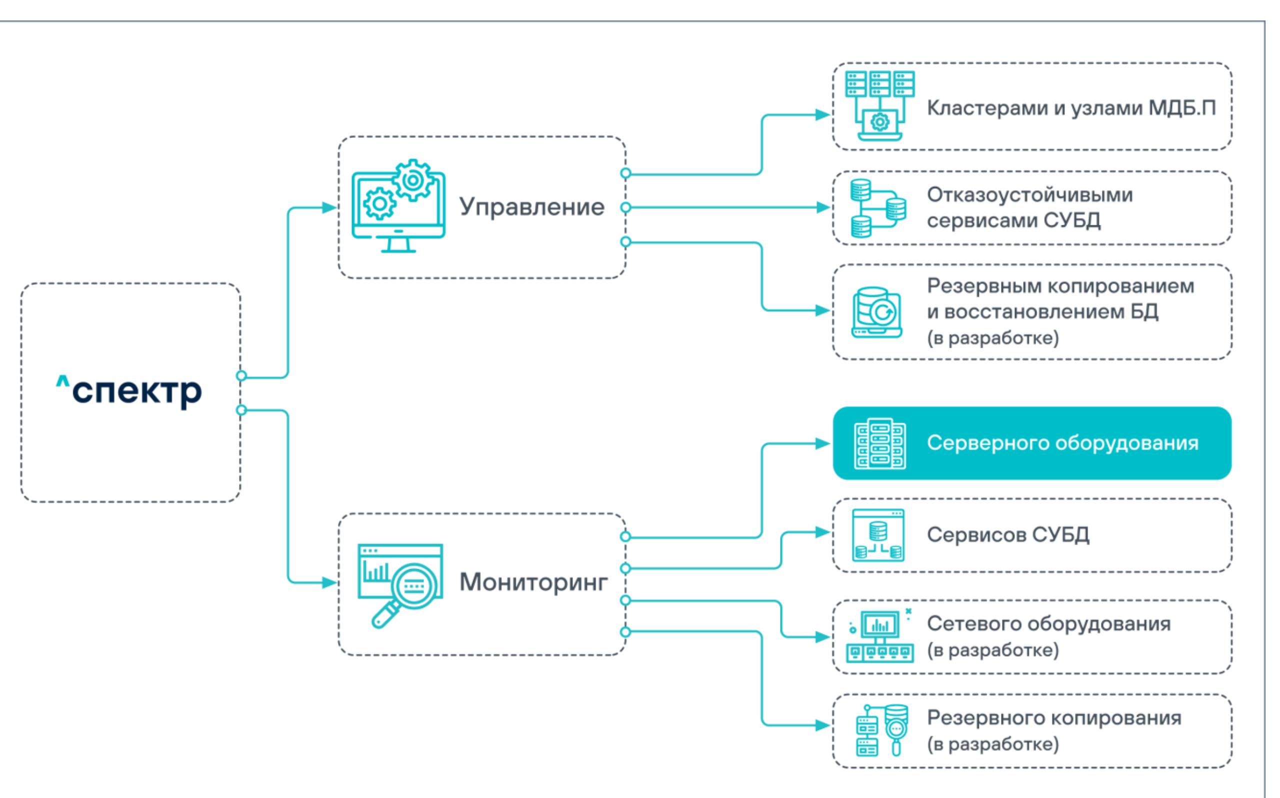
Quick CPU / Mem / Disk

ЦПУ 0.975%
Sys Load (5m ... 0.983%
Sys Load (15... 0.983%
RAM Used 5.57%
SWAP Used 0.983%
Root FS Used 32.3%
CPU C... 4
Uptime 7.2 week
RootF... 17 GiB
RAM T... 16 GiB
SWAP ... 1024 MiB

Общее о ЦПУ / ОЗУ / Сеть / Диски

CPU Basic
100%
75%
50%
25%

Memory Basic
18.6 GiB
14.0 GiB
9.31 GiB
4.66 GiB



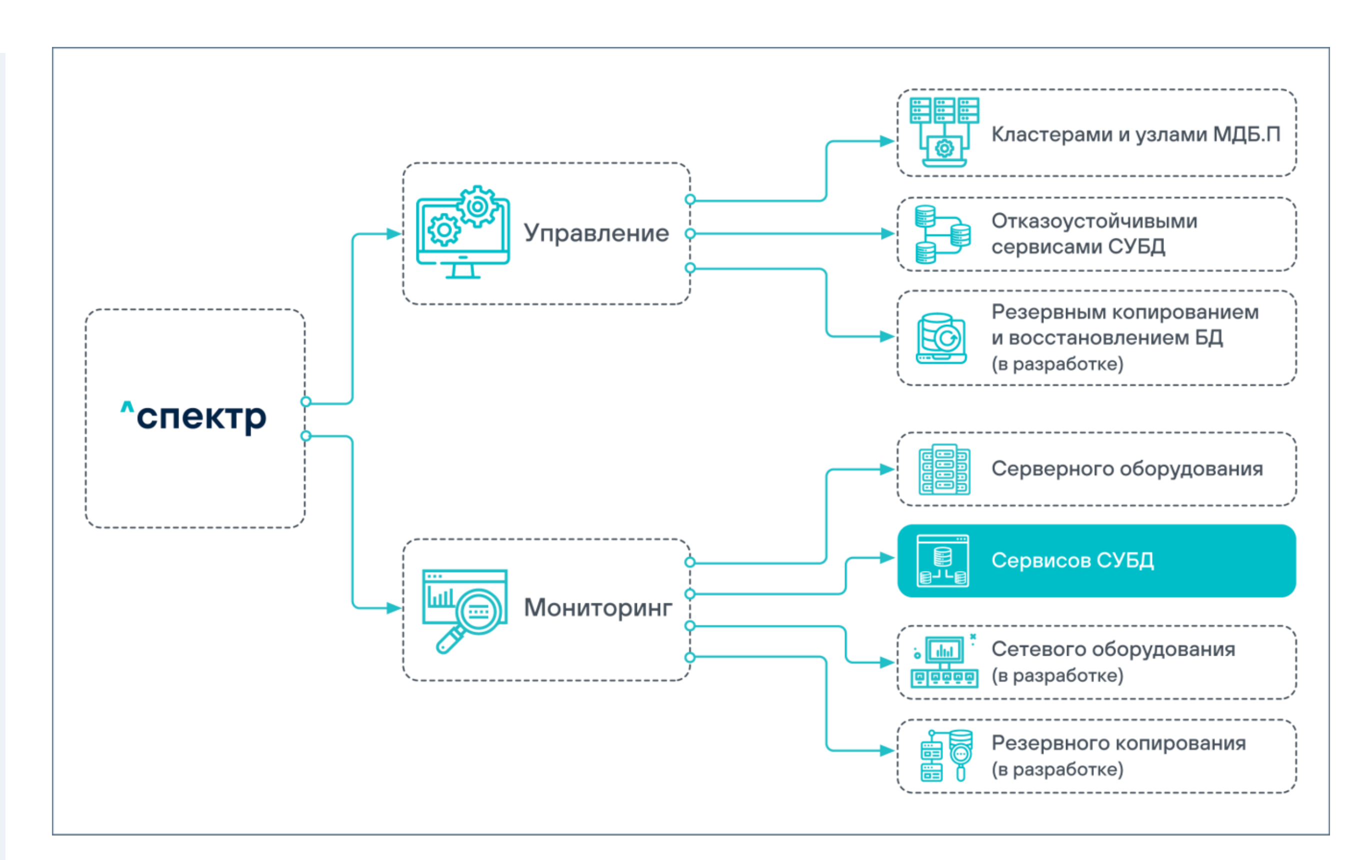


Мониторинг СУБД











Спектр. О пользе и вреде копирования



Исходное состояние системы

Состояние А → Состояние Б

↓

Состояние В → Готовая Машина

Спектр. О пользе и вреде копирования



Исходное состояние системы (?)

Состояние А

Состояние Б

Состояние Б'

Состояние Г

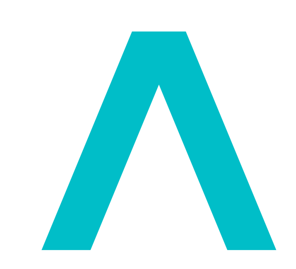
Состояние В

Состояние Д

Нет конечного состояния



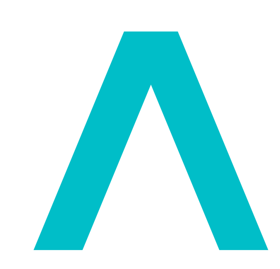
Спектр. Непрошенные советы




```
<crm_mon version="2.1.1-alt1">
  <summary>
  <nodes>
  <resources>
    <clone id="HA-1701-pg11-a9-1-clone" multi_state="true" unique="false"
    failure_ignored="false">
    <group id="HA-1701-pg11-a9-1-master-group" number_resources="2" manage
      <resource id="HA-1701-pg11-a9-1-master_ip" resource_agent="ocf:heart
      "false" blocked="false" managed="true" failed="false" failure_ignore
        <node name="1701-pg11-a9-1-01" id="1" cached="true"/>
      </resource>
      <resource id="HA-1701-pg11-a9-1-master_repl_ip" resource_agent="ocf
      orphaned="false" blocked="false" managed="true" failed="false" failu
        <node name="1701-pg11-a9-1-01" id="1" cached="true"/>
      </resource>
    </group>
    <resource id="HA-1701-pg11-a9-1-sync_ip" resource_agent="ocf:heartbea
    "false" blocked="false" managed="true" failed="false" failure_ignored=
    <resource id="HA-1701-pg11-a9-1-async_ip" resource_agent="ocf:heartbea
    "false" blocked="false" managed="true" failed="false" failure_ignored=
```




Best Practice в создании ПАК

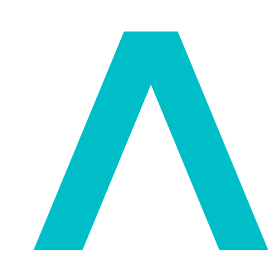


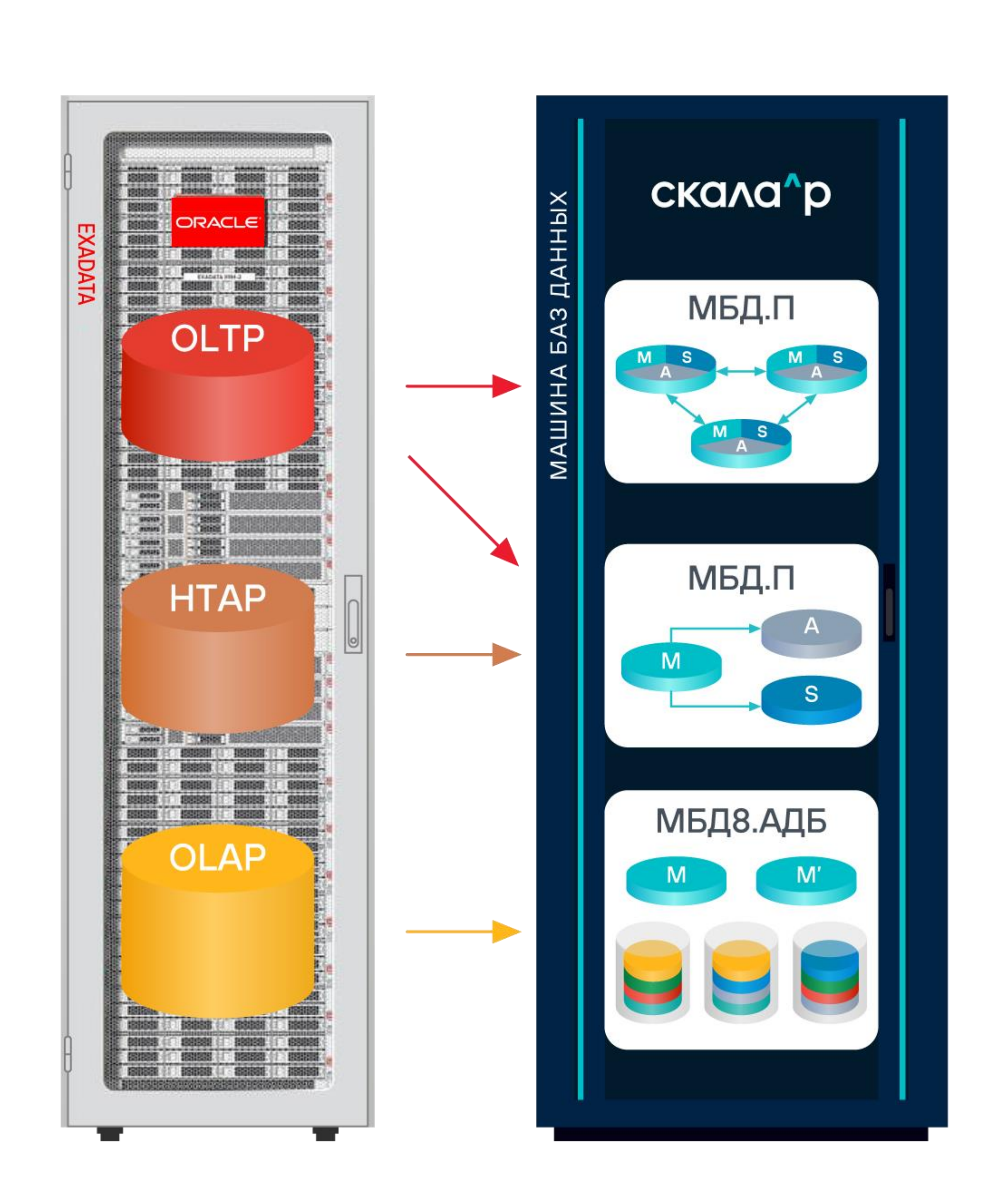
Платформенное программное обеспечение **Oracle**	Производительное оборудование
Специализированное программное обеспечение **Exadata Software**	Мониторинг и управление **Enterprise Manager**

Оптимизация



Миграция Oracle Exadata -> Скала^р МБД.П



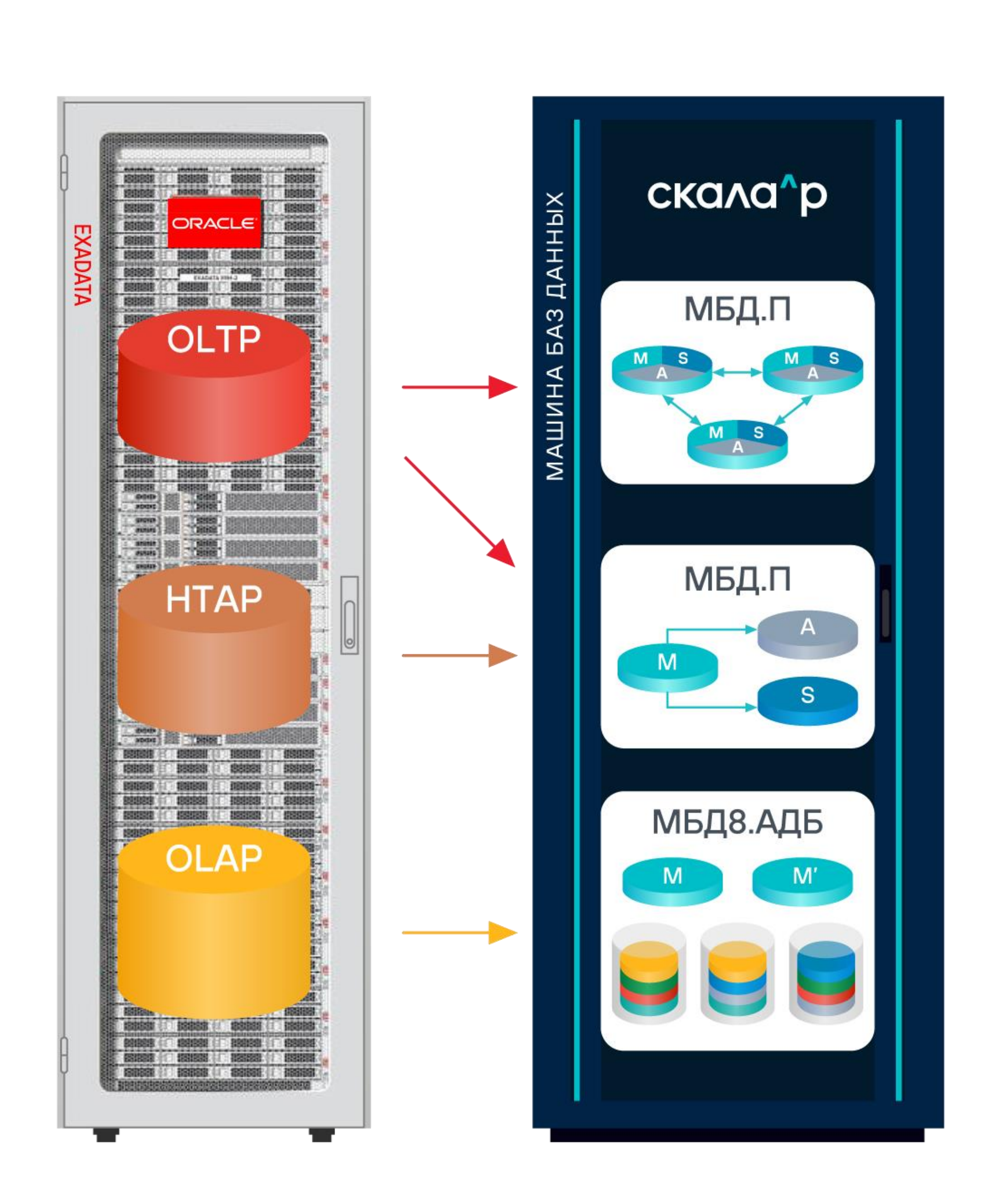


- Скала^р МБД.П в инфраструктуре заказчика
- Миграция данных (Ora2PG)
- Profit!!!



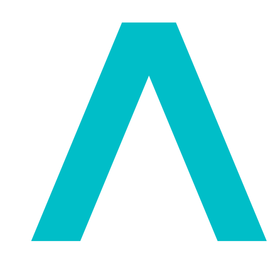
Миграция Oracle Exadata -> Скала^р МБД.П





- Скала^р МБД.П в инфраструктуре заказчика — ✔
- Миграция данных (Ora2PG) — ✔
- ~~Profit!!!~~

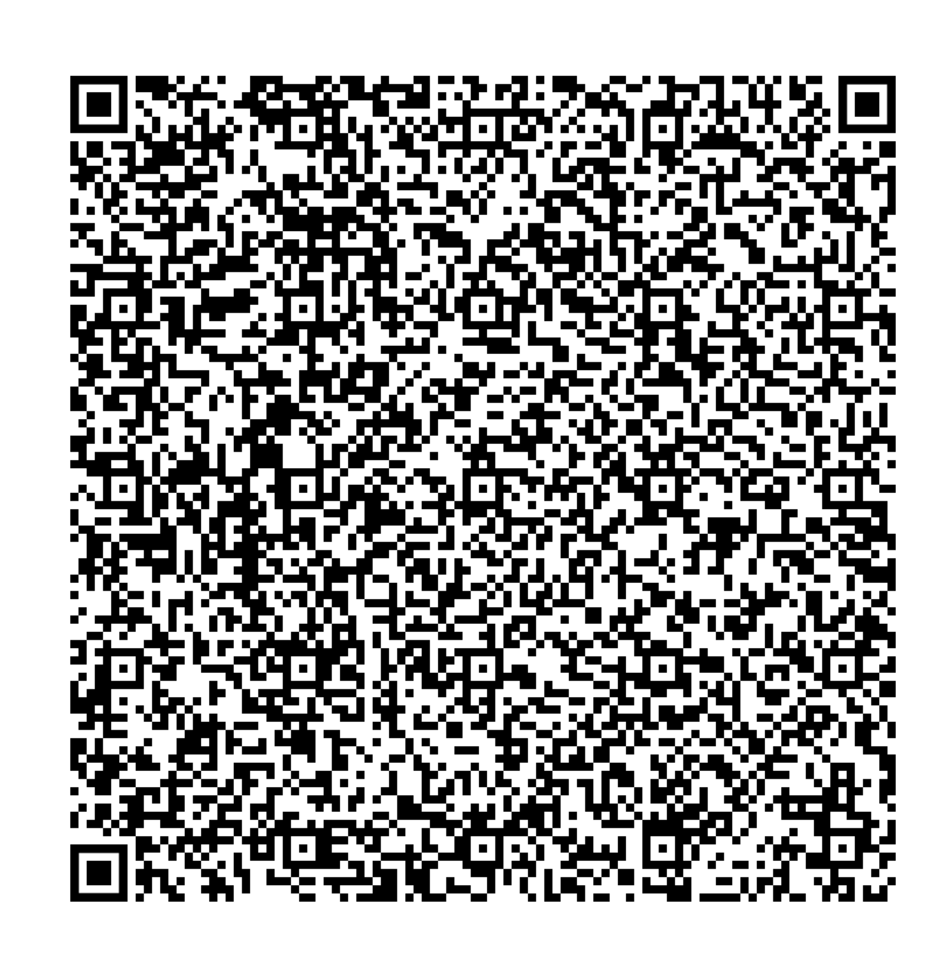
Замена СУБД в ИС



Можете привести примеры, где несложно импортозаместиться?

…

PostgreSQL, мне кажется, тоже хорошая история. Но не везде и не всегда. Oracle так просто не заменишь, но во многих случаях можно.



…А в ФНС возможно заменить?

В АИС «Налог-3» – невозможно.

Что для этого с Postgres нужно сделать?

Ничего. Невозможно.

Тотальная зависимость от Oracle?

Да, надо просто написать новую систему на другой архитектуре.

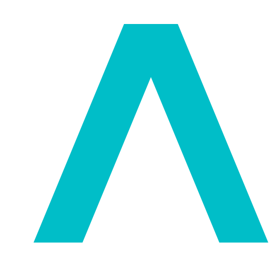
4-ю версию АИС «Налог»?

Да.





Замена СУБД в ИС

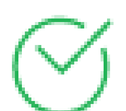


Как мы переписывали бизнес-логику высоконагруженного приложения на PLPG/SQL

Базы данных и системы хранения | Миграция / Хранимые процедуры

#Миграция / Хранимые процедуры

25 ноября, 10:00, Зал «h3: Яндекс трек»


Доклад принят в программу конференции

Мнение Программного комитета о докладе

Миллионы наших сограждан оформляют больничные, а IT-система в ФСС, которая обслуживает этот процесс, переехала из Oracle в Postgres. В докладе будет все об особенностях этого увлекательного процесса.



Скала300 -> Скала500 -> Скала700



Скала500

Скала700

Скала300



Скала300 -> Скала500 -> Скала700



Скала400

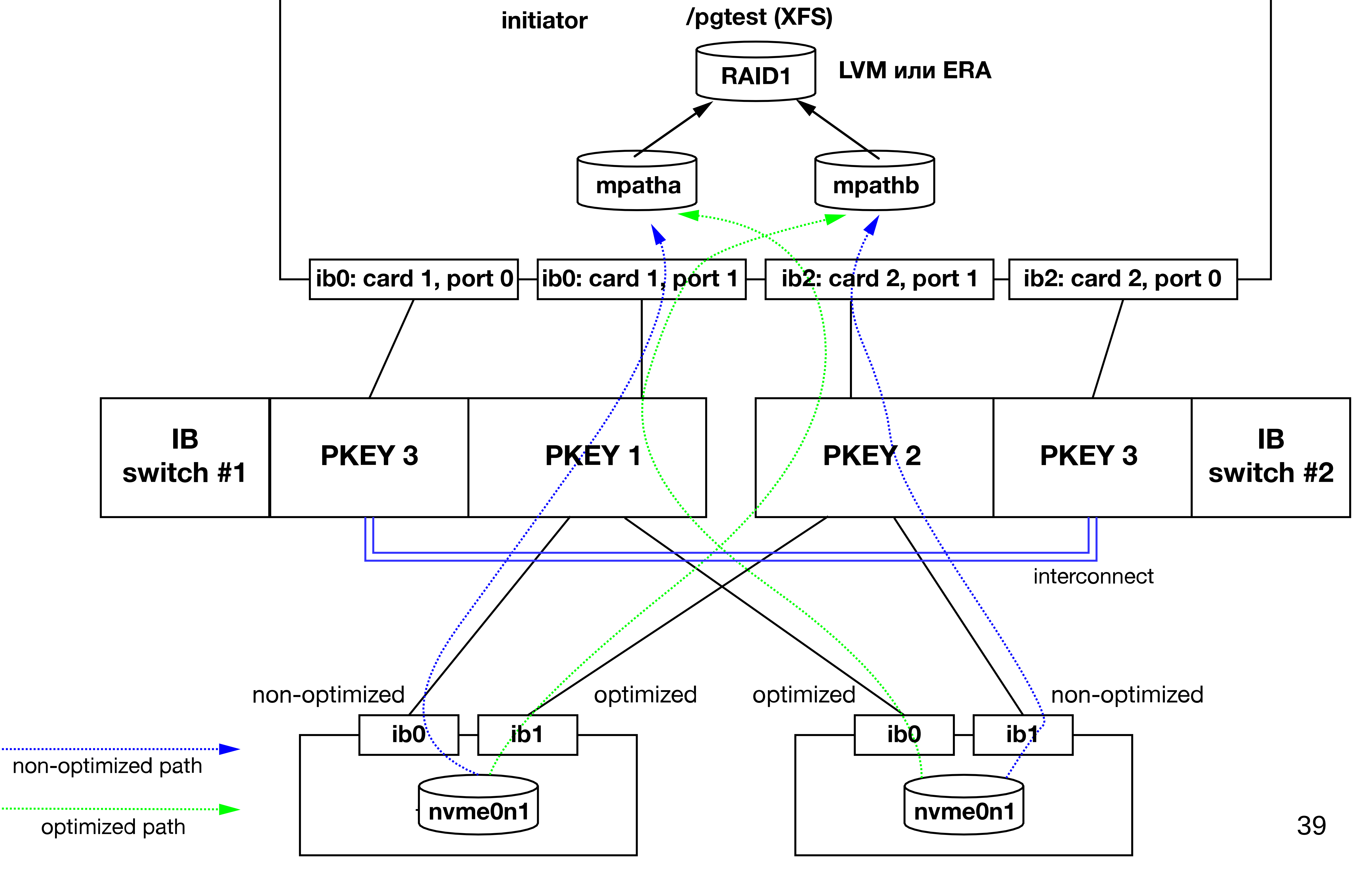
Скала700

Скала300

Скала500 NVMe over Fabric



Скала300 | **Скала500**



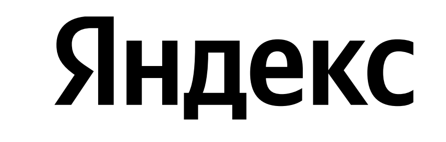




Про CPU: Xeon & Epic -> Эльбрус и Байкал -> ARM64



Архитектура	Модель	Доступность	Примечание
x86-64	Intel Xeon/AMD Epic	+/-	Полностью устраивают
VLIW	Эльбрус	😟	Производительность?
ARM64	Байкал-М	😫	Не успели попробовать
ARM64	«Китайские» ARM	+/-	Производительность?
RISC-V	Байкал-S	?	Ждём
RISC-V	Решения от Yadro	?	Ждём





Тесты на Huawei TaiShen



PGBENCH TPC-B TRANSACTIONS

HUAWEI TaiShan 200 Server (ARM)
HUAWEI FusionServer Pro (x86)

TPS: 0, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000

CONNECTIONS: 0, 50, 100, 150, 200, 250, 300, 350, 400, 450

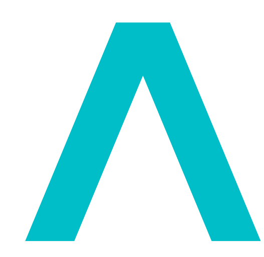
PGBENCH SELECT-ONLY TRANSACTIONS

HUAWEI TaiShan 200 Server (ARM)
HUAWEI FusionServer Pro (x86)

TPS: 0, 200000, 400000, 600000, 800000, 1000000, 1200000

CONNECTIONS: 0, 50, 100, 150, 200, 250, 300, 350, 400, 450

Тесты на Huawei TaiShen



Пример конфигурации сервера БД PostgreSQL (postgresql.conf):

```
max_connections = 1024
shared_buffers = 390GB
max_prepared_transactions = 2048
huge_pages = try
work_mem = 1GB
maintenance_work_mem = 2GB
dynamic_shared_memory_type = posix
```

Методика:

1) Создаем тестовую БД следующей командой (размер около 370 ГБ):

```
pgbench -i -s 25000
```

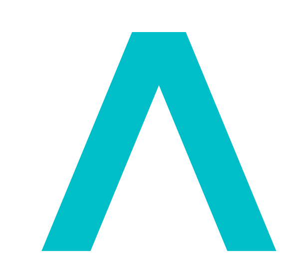
2) Выполняем несколько SQL-запросов для разогрева кэша БД:

Это не хорошо и не плохо, просто результаты нерелевантны для наших условий!!!





Тестирование TaiShen в IBS Интерлаб



tps

Подключений	50	100	200	300	400	500	600	700	800	900	1000
tps		11,389	4,159	4,750	6,163	3,284	7,381	6,511	6,273	6,187	4,449







openGauss



<table>
<tr><td colspan="3">Ответвление отPostgreSQL 9.2</td></tr>
<tr><td>C++ вместо C</td><td colspan="2">Thread-per-connection вместо process-per-connection</td></tr>
<tr><td colspan="3">Основная особенность — трёхдвижковость</td></tr>
<tr><td>Строчный</td><td>Резидентный</td><td>Столбцовый</td></tr>
<tr><td colspan="3">Специфика по работе
с NUMA (Куньлун, Куньпен)</td></tr>
<tr><td colspan="3">Модифицированное журналирование</td></tr>
<tr><td>Многопоточное применение
журнала предзаписи</td><td colspan="2">Журналы отмены (практически REDO LOG у Oracle!!!)</td></tr>
</table>

Sun
ORACLE
EXADATA
МАШИНА БАЗ ДАННЫХ
скала^р

скала^р

отсекая лишнее



Спасибо!

Оценка

Константин Аристов

Скала^р, техлид

Karistov@skala-r.ru

@A_K_M_74



Яндекс